



The Seven Steps: Building the DiGA Thesaurus

RESEARCH PAPER

]u[ubiquity press

SERENA AUTIERO

FREDERIK ELWERT

CRISTIANO MOSCATELLI

JESSIE PONS

*Author affiliations can be found in the back matter of this article

ABSTRACT

This article presents the creation of the ‘Digitization of Gandharan Artefact Thesaurus’, a digital resource for the description of Gandharan art and – more generally – of Buddhist art. It was developed by the project ‘Digitisation of Gandharan Artefacts: A project for the preservation and the study of the Buddhist art of Pakistan’ (DiGA). This work has been realized by walking Seven Steps, a serendipitous coincidence pairing our work to a famous episode in the early life of the Buddha. Our Seven Steps are:

1. Collection and selection of sources, both digital and printed.
2. Creation of the core of the Thesaurus by digitizing the *Repertorio Iconografico per la schedatura dell’arte gandharica*.
3. Modifying, enriching, and structuring the Repertorio in RDF/SKOS.
4. Implementation of a hierarchy of concepts for narrative scenes (“narratives”).
5. Implementation of a hierarchy of concepts for characters (“figures”).
6. Comparison and reconciliation with other already existing digital thesauri.
7. Collaboration with other projects and institutions.

The DiGA Thesaurus marks an important milestone towards a new vision of Gandharan studies in which the availability of digitized collections will foster new insights and understanding. Already during the creation of the Thesaurus, exciting new perspectives have opened in the field of both Gandharan studies and digital humanities.

CORRESPONDING AUTHOR:

Serena Autiero

Centre for Religious Studies,
Ruhr Universität, Bochum,
Germany

serena.autiero@rub.de

KEYWORDS:

thesaurus; Buddhist art;
Gandhara; SKOS; LOD

TO CITE THIS ARTICLE:

Autiero, S., Elwert, F.,
Moscatelli, C., & Pons, J.
(2023). The Seven Steps:
Building the DiGA Thesaurus.
*Journal of Open Humanities
Data*, 9: 11, pp. 1–14. DOI:
[https://doi.org/10.5334/
johd.111](https://doi.org/10.5334/johd.111)

(1) CONTEXT AND MOTIVATION

(1.1) INTRODUCTION

According to Buddhist tradition, immediately after his miraculous birth emerging from the right side of his mother Māyā Devī, the infant Siddhārtha Gautama stood up, took seven steps, and announced his imminent Awakening and hence, his last existence (for an overview on the hagiography of the historical Buddha with references to the primary sources see [Strong 2001](#)).

This episode, one which belongs to the Buddha's birth cycle, offers a fitting metaphor for the DiGA Thesaurus. This Thesaurus, a digital resource for the description of Buddhist art from Gandhāra, was created during the infancy of the DiGA Project (short for Digitization of Gandhāran Artefacts: A project for the preservation and the study of the Buddhist art of Pakistan). It was developed for the benefit of the scholarly community with the intent to cover gaps in the terminology available to describe our objects of study: Buddhist sculptures. The Thesaurus was recently born and unlike Siddhārtha Gautama, we doubt that this will be its final form of existence and we expect new versions to be gradually released in the future. We nevertheless wish to introduce this important milestone and the seven steps which led to its completion and hope that these will appear equally auspicious as those taken by Siddhārtha Gautama.

(1.2) THE DIGA PROJECT IN THE CONTEXT OF GANDHARAN STUDIES

DiGA is a project that aims to digitize and catalogue a corpus of almost 2000 Buddhist sculptures produced during the first centuries of the Common Era in ancient Gandhāra that are now preserved in the Dir Museum Chakdara and the Swat Museum, Saidu Sharif (Province of Khyber-Pakhtunkhwa, Pakistan). Gandhāra, a historical region which covers present-day North-western Pakistan and Eastern Afghanistan, was a pivot between South and Central Asia and, in the first centuries of the Common Era, a thriving Buddhist centre. The iconographic and formal features of the sculptures bear testimony to the rich cultural and religious heritage of this region.

The great majority of the sculptures the project digitizes has been discovered during excavations conducted by the Department of Archaeology and Museums, Khyber-Pakhtunkhwa, Government of Pakistan, the Department of Archaeology, University of Peshawar, and the Italian Archaeological Mission in Pakistan. They come from a dozen of Buddhist sites located around the modern city of Chakdara (district Dir) on the right bank of the Swat River.¹ Unlike many other collections of Buddhist art from Gandhāra, the archaeological provenance of the objects and their entry in the Dir Museum are documented. Their digitization therefore provides a solid corpus for the reassessment of crucial research questions in the field of Gandharan studies. In this respect, three concrete scientific desiderata underpin our project and the digital concept on which it relies ([Amato, Elwert, & Pons, 2022](#)). These generally pertain to the definition and development of the Gandharan artistic school and to the relationship between images and texts in the study of Buddhism in the region. Put briefly, there are:

- To give more visibility to the multiplicity of Gandharan styles by bringing to the fore certain productions characterized by figures that contrast with images displaying strong Classical lines often exhibited in permanent collections. The project thereby aspires to tell a story about Gandharan art which is alternative to the Graeco-Buddhist one.²
- To map formal and iconographic variations discernible in the motifs carved on reliefs and subsequently identify and circumscribe workshops, production centres, stylistic zones, and sub-schools. This approach may therefore cast light on the historical development of Gandharan art ([Faccenna, 2001](#)).
- To reassess visual material in light of other types of sources, primarily Buddhist texts, which may have been in circulation in the region of Gandhāra when the sculptures were produced ([Salomon, 2018](#)). This can potentially inform us on the various Buddhist traditions present in Gandhāra in general and Swat in particular.

¹ These are the sites of Chatpat, Bambolai, Ramora, Damkot, Andan-dheri, Jabagai, Macho, Tribanda, Amlok-dara, Nasafa, Shalizara, Shamsi khan, and Gumbat.

² For critical assessments of approaches to Gandharan art framed by the dichotomy between West and East see [Taddei 1980](#); [Filigenzi 2012](#); [Falser 2015](#); [Pons 2017](#).

Besides these considerations more specifically related to Gandharan studies, the DiGA Project endeavours to preserve and make widely accessible a significant facet of the cultural heritage of Pakistan. The Buddhist art of Pakistan, as other tangible and intangible assets of Pakistani cultural heritage, has incurred decades of unscientific exploration, looting and armed conflicts, resulting in its gradual dismantling, dispersal, and destruction. By digitizing these sculptures and documenting their related metadata, the DiGA Project strives to safeguard this important historical testimony. It also aims to secure the objects' long-term and widespread accessibility; the sculptures are largely unpublished as only about 10% have been reproduced in archaeological reports (Dani, 1968–69) and exhibition catalogues (Drachenfels & Luczanits, 2008). Moreover, since they are located in the provinces of Lower Dir and Swat in Pakistan, the two collections are also difficult to travel to. The digitized items are progressively imported on a database, hosted on heidICON, the multimedia platform of Heidelberg University Library, and made freely available.³ In order to comply with current standards in the field of heritage documentation, developing a standard list of terms that would allow to describe all the aspects of the corpus (e.g. information about the sculpture itself and about what it represents) was a necessity. A terminology that captures the diversity of motifs and iconographies depicted and would allow to consistently refer to and typify our objects of study also appeared to be a prerequisite to exploring the research avenues outlined above.

(1.3) THE WHAT, WHY AND HOW OF THE DIGA THESAURUS

The DiGA Project started in February 2021 and the development of the Thesaurus was one of its very first tasks. Our approach was to build on established existing work, both digital and analogue. Existing authoritative vocabularies such as the Getty Arts and Architecture Thesaurus (AAT) and IconClass are, however, limited in their treatment of non-Western iconographies. As for vocabularies developed by projects specifically dedicated to Buddhist material (e.g. Jataka Stories,⁴ Buddhist Murals of Kucha on the Northern Silk Road),⁵ they focus on types of representations which only partly find equivalents in the Gandharan visual record. While we do map these vocabularies when possible (see 3.1.6), our thesaurus primarily draws upon a limited selection of analogue sources with a focus on Gandharan Buddhist art. First and foremost among these is the *Repertorio Terminologico per la schedatura delle sculture dell'arte gandharica*, a bilingual resource in Italian and English (Faccenna & Filigenzi, 2007; see chapter 4.1 for further information on the sources).

As this article will highlight, the DiGA Thesaurus version that has been released can be fruitfully applied for cataloguing and for research purposes. It primarily intends to contribute to the development and use of a more specific terminology for Gandharan and Buddhist art and architecture, but its implications and applications are manifold:

- The DiGA Thesaurus enriches the “corpus” of existing vocabularies with one that is specific to Buddhist art. It has been originally conceived to function as a controlled vocabulary for describing objects within the DiGA Project by linking it to the heidICON database. Very concretely, the controlled vocabulary—available on SKOSMOS—was imported on heidICON as a project-specific list. As such, it complements GND (Gemeinsame Normdatei) authority data used by heidICON for cataloguing and can be used to enhance the description of the scenes depicted by adding specific keywords.
- By its open access publication as a Simple Knowledge Organization System (SKOS) resource, the DiGA Thesaurus serves as a best practice example in the field of Buddhist art through the introduction of controlled vocabularies for figures, monuments, architectural components, objects, narratives, etc.
- The DiGA Thesaurus complies with Linked Open Data (LOD) principles in an effort to bridge different collections of Gandharan Buddhist resources in different media (e.g. text and visual art) and to link resources specific to Buddhist studies to generic art historical resources such as the Getty AAT or IconClass.

³ <https://heidicon.ub.uni-heidelberg.de/> (last accessed: 23 June 2023).

⁴ Edinburgh University, direction Dr. Naomi Appleton, <https://jatakastories.div.ed.ac.uk/> (last accessed: 23 June 2023).

⁵ Saxon Academy of Sciences and Humanities, direction Dr. Monika Zin, <https://kuchatest.saw-leipzig.de/> (last accessed: 23 June 2023).

- The open nature of the whole project creates an environment that fosters collaboration through networking, the establishment of working groups and the use of services such as GitHub, encouraging scholars with similar scientific desiderata to have a source that is always available for designing similar tools.
- The use of LOD strategies to bridge Gandharan art history to more generic art historical resources will also foster a more widespread understanding of this artistic phenomenon in its own context, going beyond a still prevailing Eurocentric point of view.
- The DiGA Thesaurus also aims at mitigating the issue of using Western terminology to describe a Buddhist corpus or that of equating English and Sanskrit terms which do not overlap. This correspondence is not necessarily unambiguous, and it reflects a way of adapting a specific tradition to norms developed for another tradition.

(2) DATASET DESCRIPTION

OBJECT NAME

DiGA Thesaurus

FORMAT NAMES AND VERSIONS

Turtle (.ttl)

CREATION DATES

2021-06-16–2023-02-21

DATASET CREATORS

Serena Autiero, Data Curation, Ruhr Universität Bochum

Frederik Elwert, Conceptualization/Funding acquisition/Software, Ruhr Universität Bochum

Cristiano Moscatelli, Data Curation, Ruhr Universität Bochum

Jessie Pons, Funding acquisition/Supervision, Ruhr Universität Bochum

LANGUAGE

English, Italian, Sanskrit

LICENSE

CC0

REPOSITORY NAME

Zenodo, <https://doi.org/10.5281/zenodo.7760869>

PUBLICATION DATE

2023-03-22

(3) METHOD

(3.1) THE SEVEN STEPS OF THE DIGA THESAURUS

In this section we offer an overview of the seven steps that led to the creation of thesaurus and its very first applications, focussing in particular on the establishment of the existing set of concepts and their hierarchical organization.

(3.1.1) Step 1: Selection of sources

Gandharan art, let alone Buddhism and Buddhist art, has generated an immense bibliographical corpus and collections of Buddhist texts and images are now increasingly populating the digital space. However, these resources can be difficult to navigate. A same event in the life of the

Buddha may for instance have different labels (e.g. awakening bodhi nirvāṇa, enlightenment, all refer to the same episode). To harvest the terminology needed for our Thesaurus and feed our core vocabularies of motifs, narratives, and persons, we identified a limited set of primary sources. For the field of Gandharan studies, a vocabulary already exists: *The Repertorio Terminologico per la Schedatura delle Sculture dell'Arte Gandharica* (Faccenna & Filigenzi, 2007, hereafter *Repertorio*). It includes a wide range of topics: decorative motifs, people, fauna, flora, weapons, musical instruments, ceremonial objects, everyday objects, furniture, and means of transport. As such, it contains most of the terms needed for the basic metadata related to our sculptures as well as an important proportion of the terminology necessary to describe the scenes and figures depicted thereon. Since the scope of the *Repertorio* is very broad, but not exhaustive, we have also drawn upon other sources. These are briefly considered here in alphabetic order.⁶ The extent to which they are integrated in our Thesaurus is discussed in the corresponding sections “Narratives” and “Figures”.

1. Ali, I. & Naeem, Q. M. (Eds.) (2008). *Gandharan Sculptures in the Peshawar Museum (Life Story of Buddha)*. Mansehra: Hazara University.
This catalogue includes sculptures which compose the nucleus of the collection preserved in the Peshawar Museum focussing on the biography of the Buddha for a total of 63 hagiographic episodes (5 *jātakas* and 59 events in the last existence) and 10 generic scenes.
2. Faccenna, D. (1962). *Sculptures from the Sacred Area of Butkara I, Swat, Pakistan*. Vols. II, 2-3. ISMEO Reports and Memoirs. Rome: Istituto poligrafico dello Stato, Libreria dello Stato.
This extensive archaeological repertory of sculptures (hitherto the largest), records over 700 sculptures coming from the Swat Valley. The provenance of the majority of artefacts is known, avoiding the risk of including forgeries. It covers over thirty Buddhist narratives (mostly events from the life story of Siddhārtha Gautama) and a very wide array of generic scenes.
3. Appleton, N. (Ed.) (2019). *Jataka Stories database*. Edinburgh: Edinburgh University.
Retrieved from: <https://jatakastories.div.ed.ac.uk/> (last accessed: 23 June 2023).
This database records a total of 749 stories across eight textual collections in Indic languages as well as 122 depictions from sites in South Asia. The list of story clusters it provides is particularly valuable.
4. Jongeward, D., Lenz, T., Neelis, J. & Pons, J. (Forthcoming). *Buddhist Rebirth Narratives in Literary and Visual Cultures of Gandhāra*. Seattle: University of Washington Press.
This monograph presents a survey of past lives of the Buddha available in literary and visual forms in Gandhāra to reassess their religious and cultural significance during the early centuries CE in the region. Concerning images, it identified 15 *jātakas* carved on a total of about 176 reliefs or paintings.
5. Pons, J. (2011). *Inventaire et Étude Systématiques Des Sites et Des Sculptures Bouddhiques Du Gandhāra: Ateliers, Centres de Productions*. [Unpublished PhD thesis]. Paris-Sorbonne University.
This PhD dissertation, dedicated to the stylistic study of Buddhist sculptures from Gandhāra, compiles a corpus of 3522 provenanced sculptures excavated at 69 archaeological sites. The terminology for narratives (ca. 120 labels for hagiographical events and 70 for generic scenes) and figures (ca. 70 terms) are partly adopted from A. Foucher's seminal work *L'art gréco-bouddhique du Gandhāra: étude sur les origines de l'influence classique dans l'art bouddhique de l'Inde et de l'Extrême-Orient* (Foucher, 1905-1951). They have been translated into English for the preparation of the DiGA Thesaurus.
6. Zwalf, W. (1996). *A Catalogue of Gandhara Sculptures in the British Museum*. 2 vols. London: British Museum.
This catalogue documents 670 sculptures. Over two hundred of these illustrate the life story of Siddhārtha Gautama and *jātakas* altogether covering about forty events.

⁶ This list is largely reproduced from Elwert and Pons 2020. When relevant, it has been updated with new data.

(3.1.2) Step 2: Digitizing the Repertorio

Since the *Repertorio* encompasses much of what we envisioned for the DiGA Thesaurus and is an established resource, we aimed to convert it into an electronic resource. The *Repertorio* has a hierarchical structure which translates visually clearly in the typography and layout. It is divided into twelve parts or categories of topics. Each part is itself divided into sections and subsections. Their hierarchical relation is signalled by distinct font sizes. Each page (or “Plate”) of the *Repertorio* displays a tabular structure with three columns, from the left: illustrations, terms (or “lemmata”) in Italian and in English. In addition to the three columns, one can also distinguish rows in the page layout, resulting in a table-like structure. The generic terms in turn can have sub-entries that are views or parts of a generic entry (“secondary lemmata indicating components or specific types”). These sub-entries take the form of a hierarchical list, with numerical list entries for generic terms, and alphabetical entries for object parts (see Figure 1). Early in the project we envisioned that it should be feasible to rely on these features to re-create the hierarchical structure in a digital form.

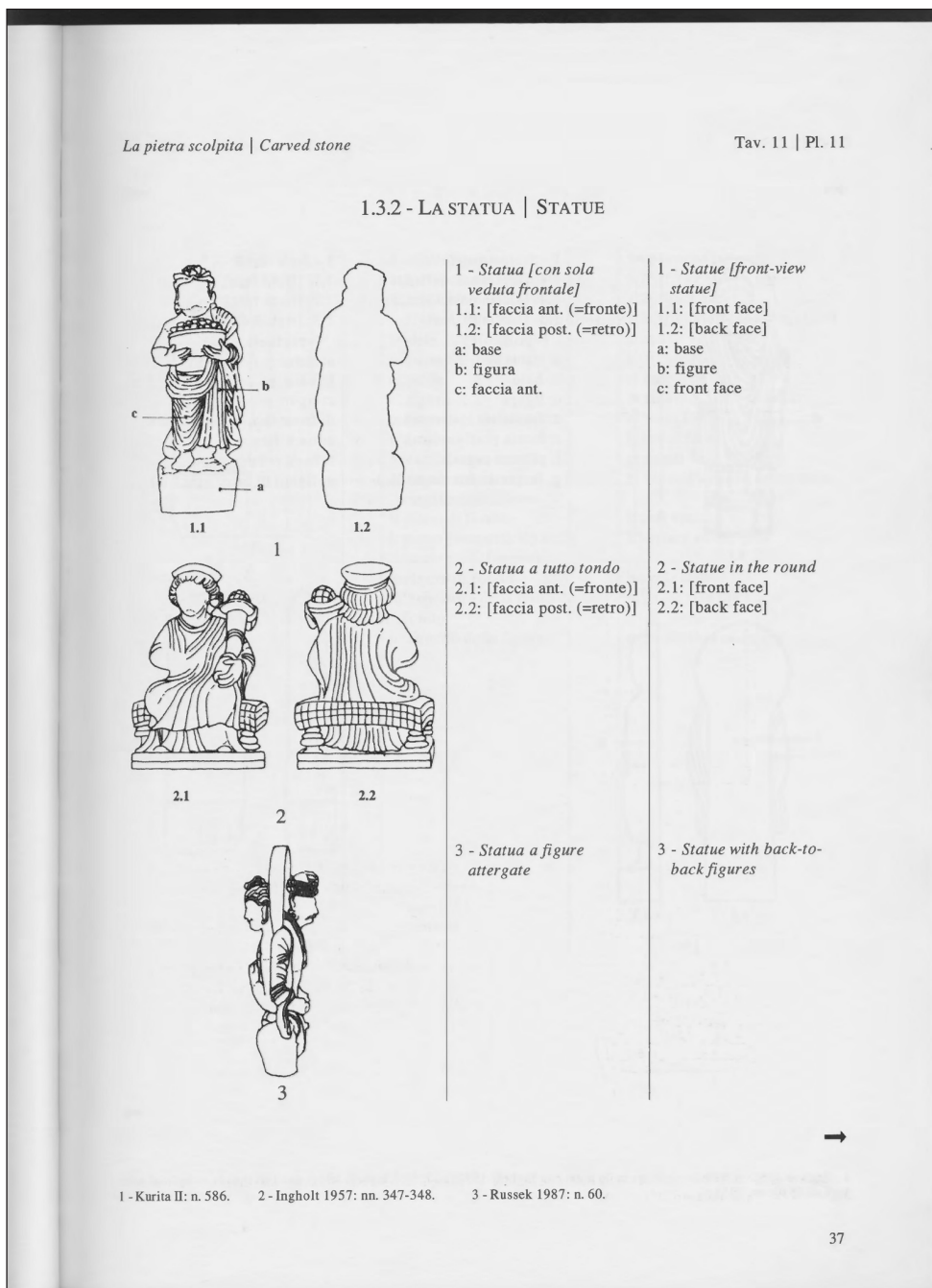


Figure 1 An exemplary page from the *Repertorio*.

In a first approach, we evaluated existing OCR and layout recognition systems. Table parsing is still a difficulty for OCR tools. The page layout of the *Repertorio* can be read like a table, but it is typographically not easily recognizable. In the end, we achieved the best results with the

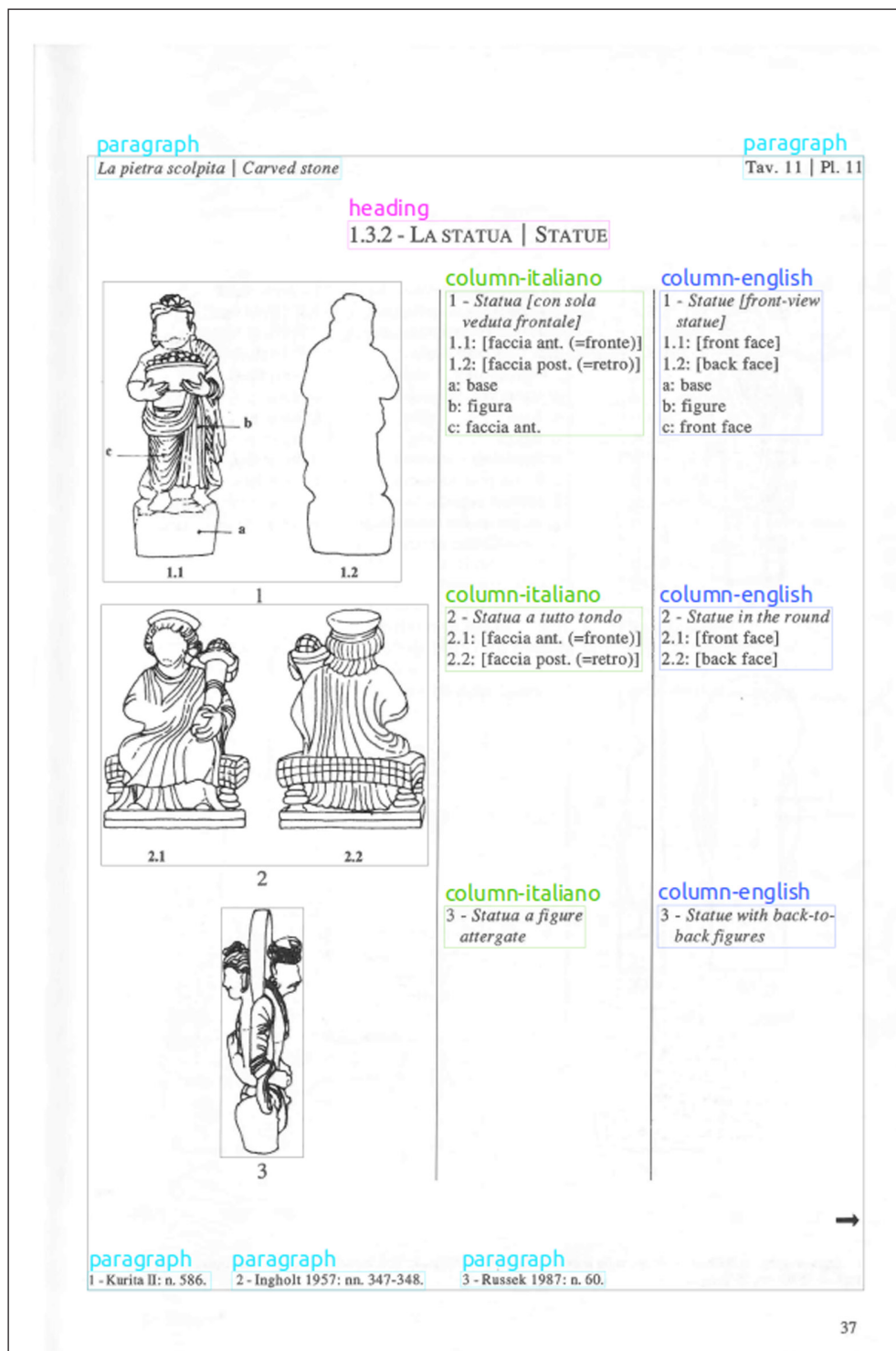


Figure 2 A page of the Repertorio after automatic layout recognition.

We could then run a default OCR model (Transkribus print 0.3) with good success. Corrections were mainly required for some special characters which served an important role in the later post-processing. The results were then exported as PAGE XML files for post-processing.

The aim of the post-processing was to re-create a machine-readable hierarchical structure from the page layout. Given the layout information in the PAGE XML files, we could employ a simple heuristic to recognize the table-like structure: each row of the table would comprise an illustration in the first cell, then a cell in Italian, and then a cell in English. Inside the cells, we leveraged regular expressions to extract the structure of the lists. Using this information,

(3.1.3) Step 3: Building a controlled vocabulary

A machine readable *Repertorio* was only the first step towards a thesaurus. To be re-usable and to comply with interoperability requirements, we converted the digital version of the *Repertorio* into the Simple Knowledge Organization System (SKOS) data model (Miles & Bechhofer, 2009). SKOS provides basic classes and properties to describe controlled vocabularies, thesauri, and similar data structures. It can easily be extended through other RDF properties. In specifying our data model, we drew inspiration from the semantic representation of the AAT (The J. Paul Getty Trust, 2014). In addition to the base SKOS classes and properties, we use the following extensions:

We use `dct:source` to specify the resource whence we derived each concept.

We use `dc:identifier` to specify the numerical ID (which is also part of the URI), following AAT's model.

We make use of the SKOS eXtension for Labels (SKOS-XL). This complicates the structure as labels are not simply literals, but instances of their own. Thus, each label, preferred or alternative, in each language has its own unique identifier. This allows us to specify not only the source of the concept itself, but also the source of a label. This is relevant as we incorporate additional labels and/or translations from other sources.

Instead of the generic `skos:broader` to construct the term hierarchy, we make use of the more nuanced relations from `skos-thes`. This allows us to differentiate the relation between a specific concept and a more generic one (`skos-thes:broaderGeneric`) from a whole-part-relation (`skos-thes:broaderPartitive`). This is a distinction that the *Repertorio* makes consistently, and which we wanted to maintain.

We use `foaf:depiction` to refer to each concept's corresponding illustration from the *Repertorio*.

For a very basic example of an entry from the SKOSified *Repertorio*, see Figure 3.

```
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix dct: <http://purl.org/dc/terms/> .
@prefix diga_source: <https://w3id.org/diga/source/> .
@prefix diga_terms: <https://w3id.org/diga/terms/> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix skosxl: <http://www.w3.org/2008/05/skos-xl#> .

<https://w3id.org/diga/terms> a skos:ConceptScheme .

diga_terms:2838159259 a skos:Concept ;
  dc:identifier "2838159259" ;
  dct:source diga_source:repertorio ;
  skos:topConceptOf <https://w3id.org/diga/terms> ;
  skosxl:prefLabel diga_terms:1135055502,
    diga_terms:1266854144 .

diga_terms:1135055502 a skosxl:Label ;
  dct:source diga_source:repertorio ;
  skosxl:literalForm "tools for working stone"@en .

diga_terms:1266854144 a skosxl:Label ;
  dct:source diga_source:repertorio ;
  skosxl:literalForm "strumenti per la lavorazione della pietra"@it .
```

Figure 3 A rendering of a DiGA Thesaurus entry in Turtle syntax.

⁷ The full, commented code is available at https://github.com/DiGArtefacts/repertorio/blob/main/Extract_OCR.ipynb (last accessed: 23 June 2023).

For the URIs used as identifiers of concepts, we are using the service w3id.org, which enables to register custom URI prefixes at no cost. URIs are implemented as redirects to URLs where the actual resources are provided. For our purposes, this offered a good balance between issuing our own URIs (and hence being responsible for their long-term availability) and using costly institutional URI services.

Checking all the circa 3000 entries imported from the *Repertorio* and implementing the necessary changes to the hierarchy proved to be time-consuming. Having done so, however, guarantees that we can offer a resource that is reliable, sustainable, and which can accommodate gradual extensions.

(3.1.4) Step 4: Extending the thesaurus: Narratives

To cover the full range of types of depictions found in the Gandharan visual record, the Thesaurus includes two main expansions to the original set of concepts of the *Repertorio*: “Narratives” and “Figures”. The “Narratives” section includes short descriptors that unequivocally identify scenes documented in Gandharan art. The DiGA Thesaurus includes three types of narratives:

1. The last existence (ca. 135 entries).
2. Previous births (*jātakas*). This grouping includes fifteen stories depicted in Gandharan art, identified according to the terminology laid out in Jongeward et al. (forthcoming).
3. Generic scenes. These show hunters, devotees, revelry scenes, worshippers and the like (ca. 50 entries). Some of these might refer to stories which have until now defied any identification attempt.

Paying attention to the actions performed potentially adds to our understanding of unidentified scenes. Usually, certain configurations of figures, motifs, or objects can be compared to known stories to facilitate their identification. Thus, looking at narratives in the perspective of developing an annotation schema would ideally allow to identify narratives through their constituent elements (Elwert & Pons, 2020, p. 9).

Once the types of narratives to include in the Thesaurus were identified, it was necessary to organize the related concepts in a hierarchy that would make the data coherent and univocal. Indeed, the category “narratives” hides a variety of concepts that needed a careful reflection to derive a functional hierarchy that could convey the complexities while remaining simple and intuitive, therefore suitable for SKOS design. Using a hierarchy for previous births was superfluous since the amount of data is limited to fifteen concepts but it was imperative for the numerous episodes from the last existence of Siddhārtha Gautama. While no consensus exists in scholarship, Ali and Qazi (2008) propose a system that is intuitive and adequate. Based on the rich collection of Gandharan art of the Peshawar Museum, the authors divide the hagiography of the Buddha in cycles (e.g. Birth cycle Princely life). From this subdivision we derived the five broader concepts of our hierarchy of the “last existence”. Naturally, the branches of the hierarchical tree are not uniform or symmetrical. Certain concepts (e.g. princely life) include up to nine narrower concepts (e.g. competitions) which in turn may include several narrower concepts (e.g. tug of war). Most of the branches end earlier however, with less fine-grained concepts. For now, the episodes in the Thesaurus are not in a chronological order, but in an alphabetical one, following standard practice for displaying SKOS vocabularies. A chronological order can be eventually implemented assigning sequential numbers to the concepts as “notations”. However, in the hagiography of the Buddha, there is not always consensus on the chronological sequence, making this endeavour problematic.

The second group of narrative concepts that needed to be systematized is “generic scenes”. In this case, hierarchy is not modelled after the sequence of episodes in the life of Siddhārtha Gautama, but is organized as a typology based on themes depicted (e.g. scenes of adoration) and the presence of certain protagonists (e.g. generic scenes with bodhisattvas). In this section we have three levels of hierarchy, with only one exception (namely “hunting scenes” having one further narrower level with two concepts).⁸

⁸ <https://w3id.org/diga/terms/49457987> (last accessed: 23 June 2023).

(3.1.5) Step 5: Extending the thesaurus: Figures

The *Repertorio* includes a Part titled “People” (Part 4) which constitutes the basic pool of concepts dedicated to protagonists carved on sculptures and their general features now integrated in our Thesaurus. The nature of Part 4 “People” and the organization of different sections, subsections and lemmata, however, resulted in an overwhelming mass of concepts once converted into SKOS. Indeed, since in the *Repertorio* concepts are illustrated referring to specific sculptural examples, extracting the data automatically generated a vast quantity of duplicates and a very uneven hierarchy. This section consequently needed a dramatic restructuring to eliminate duplicates. Moreover, several new concepts had to be derived from Jessie Pons’ previous work (2011) and added to the section. These imported concepts as well as new ones had to be organized in a coherent hierarchy to facilitate filing and browsing. We drew inspiration from the division of the *Repertorio* but extracted the section “Figures: some general features” from the other sections of “Part 4 People” which are exclusively dedicated to types of characters that appear in Gandharan narratives.⁹ “Figures: some general features” now constitutes a separate broader category within the DiGA Thesaurus. The other sections in “Part 4 People” as well as newly created concepts fed into the following types of figures which we have differentiated in the following categories:

- generic deities, spirits, mythological figures (e.g. demon, *yakṣa*)
- generic persons (including professions or roles)
- historical persons
- literary persons

It is worth noting that currently there are no historical persons listed in the Thesaurus since none can be convincingly identified in Gandharan art so far. We nevertheless decided to keep the option open.

Since digital classification is a novelty for this field, we decided to start by collecting the existing terminology in use. In cases of disagreement on the specific vocabulary it is advisable to include the existing alternatives, and then – once the dataset is more conspicuous – its very use would concur to the establishment of a standard.

The group “literary persons” branches into four parts (in alphabetical order): “bodhisattva”, “buddha”, “deities, spirits, mythological figures” and “human beings”. The general rule of separating the general features of figures from the lists of persons is not followed in the case of “bodhisattva” and “buddha”. Indeed, considering the peculiar nature of these beings, we deemed necessary to group “bodhisattva: general features” and “buddha: general features” under “bodhisattva” and “buddha” respectively.

The DiGA Thesaurus is not a closed canon and will continue to grow. Possible extensions can derive mostly from two channels: cooperation with other projects (see step 7) or progress in knowledge. It is indeed foreseeable that new iconographies requiring new concepts will emerge as more hitherto unpublished artefacts are digitized, with the increasing digitization of Buddhist art the number of necessary concepts is bound to grow.

(3.1.6) Step 6: Reconciling against existing vocabularies

The DiGA Thesaurus is built upon established printed resources for Gandharan studies and provides a machine-readable vocabulary with stable URIs for relevant concepts. This opens up new possibilities to bridge Gandharan and adjacent collections through the use of a shared reference system. In an effort to make our dataset interoperable, we seek to reconcile the DiGA Thesaurus with existing digital resources.

As indicated at the outset, we map our concepts with those used by cognate projects, such as Jataka Stories and Buddhist Murals of Kucha on the Northern Silk Road, particularly relevant for the description of previous birth stories and, to some extent, for the description of the last existence of the Buddha. The terms are documented in the DiGA Thesaurus as alternatives to the preferred terms that we used for a similar concept. The DiGA Thesaurus links back to the original resource through the URIs.

⁹ These are: 4.2 Buddha; 4.3 Bodhisattva; 4.4 Ascetic, Brahmin; 4.5 Monk, Nun; 4.6 Male figure; 4.7 Warrior; 4.8 Hunter, 4.9 Armed horseman, 4.10 Female figure, 4.10 Divine, semi-divine figure.

We also align generic concepts (e.g. animals, weapons, tools) to authoritative vocabularies, especially the AAT. Reconciling the DiGA Thesaurus against the AAT has been possible through a semi-automatic process via OpenRefine.¹⁰ The process yielded 333 automatic matches, which have been manually verified and amended, resulting in over 350 matches (as the vocabulary grows, also the number of matches is growing). Matching AAT concepts are connected to DiGA concepts through a `skos:closeMatch` relation.¹¹ A similar process can and will be applied in the future to other Thesauri such as IconClass, adding to the dynamic nature of the DiGA Thesaurus. Using these established resources for reconciliation can also mitigate the issue of having to create numerous cross-links between a growing number of project specific thesauri. The AAT can then also act as a hub to find matches between all thesauri that link to it.

(3.1.7) Step 7: Opening up the Thesaurus

The seventh and last step does not mark the end of the Thesaurus' progress. During its genesis, the Thesaurus was envisioned as a collaborative enterprise. Its development is therefore rooted in Open Data strategies to ensure interoperability and a wide accessibility.

With this vision in mind, we established a scholarly network with the purpose to share the Thesaurus with our peers and discuss with its features, its usage, and potential challenges, content-related or technical. To cater to different communities of users, the DiGA Thesaurus also exists as an open access repository on GitHub and Zenodo, making it widely accessible to the scholarly community, and facilitating comments, requests, modifications and discussions.¹² The scholarly network for the DiGA Thesaurus includes a multidisciplinary audience and offers an opportunity for open collaboration on several levels. Thus, parallel to our effort to incorporate other projects' terminology, other projects have conversely used the DiGA Thesaurus. For example, the "Upper Indus Petroglyphs and Inscriptions in Northern Pakistan" Project, led by Jason Neelis at Wilfrid Laurier University, is currently testing the DiGA Thesaurus on a corpus of Buddhist representations. These are very different in nature from the Buddhist sculptures from which our Thesaurus stems, but they remain closely affiliated from both an iconographic and technical point of view. This limited corpus of Buddhist petroglyphs encompasses an iconographic repertoire which largely overlaps with our DiGA Thesaurus which also includes concepts relevant to the technical production of petroglyphs (e.g. section "the sculptor's work" → tools for working stone). Another project led by Roy Tzohar at Tel Aviv University applies the Thesaurus for semantic tagging of the *Buddhacarita* by Āśvaghōṣa, a Sanskrit epic about the Buddha's life. This collaborative endeavour is currently (as of June 2023) feeding the Thesaurus with several new concepts, while also paving the way for a multimedia platform to explore the connection between literary traditions and iconography.

The synergy which results from a shared interest in LOD strategies and a concerted effort to use a consensual controlled vocabulary has exciting applications for research. By collectively pondering technical solutions that can enhance the analysis of our respective datasets and facilitate their comparisons across corpora in different media, we may collectively answer questions about the history of Buddhism, its local formations and spread, as well as the interplay between texts and images beyond the bounds of the region of Gandhara.

(4) RESULTS AND DISCUSSION

Constructing the Thesaurus starting from a printed source poses many challenges. This process took several months after the *Repertorio* was OCR'd.

The printed and the digital media afford different utilizations and the structure of the printed reference and that of a digital resource are not identical. The printed format requires a lot of duplication where the same generic terms can be used several times to refer to a specific component of various items. In other cases, the identical concept would appear multiple times because it belongs to multiple sections in the *Repertorio*. In that case, we deleted

¹⁰ <https://openrefine.org/> (last accessed: 23 June 2023).

¹¹ E.g. animals: <https://w3id.org/diga/terms/406891803> (last accessed: 23 June 2023).

¹² See note 5.

duplicate entries and assigned multiple broader entries (`skos-thes:broaderGeneric` or `skos-thes:broaderPartitive`) to the remaining concept, turning our structure from a hierarchical tree into a polyhierarchy.

Additionally, since certain types of garments (e.g. *uttariya*) can be characteristically worn by several figures, we decided to create a separate top-level hierarchy “figure: some general features” (see also Step 5). Entries like body parts, dresses or hair types were moved here and then linked back to the actual figures displaying them through a `skos:related` property when necessary.

This process is made necessary by the very nature of the *Repertorio* and how it was conceived. Indeed, there is a radical conceptual difference between our Thesaurus and the *Repertorio*. Our data set is based on relations, roles, and functions, while the *Repertorio* completely relies on visual documentation of existing artefacts. On the one hand, we wanted to have a more generic vocabulary, not exclusively linked to attested iconography, leaving room for expansions and new applications. On the other hand, our goal has always been to have a thesaurus that is also functional, that gives us the possibility to also describe actions and stories represented in narrative reliefs. This latter approach was pivotal in determining the idea of creating a new vocabulary for narratives.

Another important point that generated discussion is language diversity. The *Repertorio* is in English and Italian, and it has already been translated in Chinese (Faccenna, Filigenzi & Vignato, 2014). Across the text some lemmata are expressed in Sanskrit. Since the *Repertorio* was conceived as a bilingual resource, Sanskrit labels were initially listed as English or Italian and imported either as `skosxl:prefLabel`, or as `skosxl:altLabel`. Although only a minority of concepts has Sanskrit labels at the moment, we implemented Sanskrit in Latin script as a third language option, removing these inaccuracies and avoiding potential confusion. Besides offering more information and details, having Sanskrit in the language pool allows to:

- provide more detail and information on translations and correspondences of terms,
- accommodate the will and propensity of scholars to use Indic labels,
- improve the translation of discipline-specific terminology,
- minimize the use of West-centric terminologies in the description of Gandharan iconography.

Of course, introducing Sanskrit (or other languages such as Tibetan) as a separate language option implies a wider reflection, for instance about the adequacy of this language for the objects it designates in Gandhāra or elsewhere in the Buddhist world or about semantic shifts in the meaning of a word in a language and its translation in another. Further research about emic terminologies relating to Buddhist art and architecture is necessary but unfortunately beyond the scope of our current work. Nevertheless, we hope to have already started the discussion and offered a tool that will open up to progress in the field.

In general, the issues we encountered while creating the DiGA Thesaurus and the solutions we envisioned generated fruitful discussions and improved our approach to terminology and classification. Examining iconography through the classificatory prism allowed us to spot variety and transformation. This approach stimulated deeper thoughts on iconographic representations, revealed new visual features and – by pinpointing differences – made patterns more visible.

(5) IMPLICATIONS/APPLICATIONS

We are confident that with the release of the DiGA Thesaurus we accomplished the first (seven) steps to achieve an important goal and vision for a future of Gandharan studies where the availability of digital collections will lead to new insights. Already during the implementation of the Thesaurus, we envisaged interesting avenues for future research both on the side of Gandharan studies and on that of Digital Humanities.

We should perhaps indicate that the seven steps elaborated here are not all one-time endeavours, or, so to say, boxes to check. Steps 3 and 7 are, by nature, open processes that

accompany the life and function of the Thesaurus on the long run. Also Step 6, reconciling the entries with existing major vocabularies, is a process bound to be repeated and improved over the life and usage of the Thesaurus.

The DiGA Thesaurus is not a finished product. In this phase of the DiGA Project, one of the main tasks is the description of Gandharan artefacts stored in the Dir Museum (Chakdara); this activity already resulted in the integration of new terms to the Thesaurus, and the list is constantly growing.

In addition to already established collaborations,¹³ we aim to further expand the Thesaurus, the scope of which is currently on South and Central Asia, while it would be desirable to include concepts from other regions and traditions (e.g. East Asia). The DiGA Thesaurus is meant to grow and connect with similar initiatives, aiming at an increasingly comprehensive dataset for the description of Gandharan art and – more broadly – of Buddhist art.

ACKNOWLEDGEMENTS

We acknowledge support by the Open Access Publication Funds of the Ruhr-Universität Bochum.

We would like to express our sincere gratitude to Sarah Rautert (Student assistant, Ruhr Universität Bochum) for her invaluable assistance during all the phases of the DiGA Project. Her dedication, attention to detail, and hard work were instrumental in the successful completion of this project.

FUNDING INFORMATION

The project on which this article is based is funded by the Federal Ministry of Education and Research under the funding code 01UG2048X, funding line eHeritage. The responsibility for the content of this publication lies with the authors. The project is embedded in the Centre for Religious Studies, Ruhr-Universität Bochum.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS


Serena Autiero: conceptualization, data curation, resources, investigation, writing – original draft, writing – review and editing, supervision


Frederik Elwert: software, funding acquisition, methodology, writing – original draft, writing – review and editing

Cristiano Moscatelli: data curation, resources, writing – review and editing

Jessie Pons: funding acquisition, methodology, project administration, writing – original draft, writing – review and editing

AUTHOR AFFILIATIONS

Serena Autiero  orcid.org/0000-0003-1118-8910
Centre for Religious Studies, Ruhr Universität, Bochum, Germany

Frederik Elwert  orcid.org/0000-0001-9149-9377
Centre for Religious Studies, Ruhr Universität, Bochum, Germany

Cristiano Moscatelli  orcid.org/0000-0003-1043-9159
Centre for Religious Studies, Ruhr Universität, Bochum, Germany

Jessie Pons  orcid.org/0000-0003-1211-2213
Centre for Religious Studies, Ruhr Universität, Bochum, Germany

- Ali, I., & Qazi, M. N.** (2008). *Gandharan Sculptures in the Peshawar Museum (Life Story of Buddha)*. Mansehra: Hazara University.
- Amato, A., Elwert, F., & Pons, J.** (2022). *Digitization of Gandharan Artefacts: A Project for the Preservation and the Study of the Buddhist Art of Pakistan. A Digitization Concept*. Bochum: Ruhr-Universität Bochum. DOI: <https://doi.org/10.46586/rub.204>
- Appleton, N., & Harvey, P.** (Eds.). (2019). *Buddhist Path, Buddhist Teachings: Studies in Memory of L.S. Cousins*. Equinox Publishers. Retrieved from <https://www.equinoxpub.com/home/buddhist-path/>
- Dani, A. H.** (1969). Excavation at Andhandheri/Ramora/Bambolai/Chatpat/Damkot. *Ancient Pakistan*, 4, 33–151.
- Drachenfels, D., & Luczanits, C.** (Eds.). (2008). *Gandhara, The Buddhist Heritage of Pakistan, Legends, Monasteries, and Paradise*. Mayence: Philipp von Zabern.
- Elwert, F., & Pons, J.** (2020). *Linked Data Methodologies in Gandhāran Buddhist Art and Texts*. Pelagios Working Group Final Report. DOI: <https://doi.org/10.13154/rub.148.125>
- Faccenna, D.** (1962). *Sculptures from the Sacred Area of Butkara I, Swat, Pakistan* (Vols. II, 2-3). ISMEO Reports and Memoirs. Rome: Istituto poligrafico dello Stato, Libreria dello Stato.
- Faccenna, D.** (2001). Il fregio figurato dello stupa principale nell'arte sacra buddhista di Saidu Sharif I (Swāt, Pakistan). *IsIAO Reports and Memoirs*, 28. Roma: IsIAO.
- Faccenna, D., Filigenzi, A., & Istituto italiano per l'Africa e l'Oriente.** (2007). *Repertorio terminologico per la schedatura delle sculture dell'arte gandharica: Sulla base dei materiali provenienti dagli scavi della Missione archeologica italiana dell'IsIAO nello Swat, Pakistan*. Roma: IsIAO.
- Faccenna, D., Filigenzi, A., & Vignato, G.** (2014). *Repertory of terms for cataloguing gandharan sculptures: based on materials from the IsIAO Italian archaeological mission in Swat Pakistan* [犍陀罗石刻术语分类汇编: 以意大利亚非研究院巴基斯坦瓦特考古项目所出资料为基础]. Shanghai: Ancient Books Publishing House.
- Falser, M.** (2015). The Graeco-Buddhist Style of Gandhara. — A “Storia ideologica”, or: How a Discourse Makes a Global History of Art. *Journal of Art Historiography*, 13, 1–52.
- Filigenzi, A.** (2012). Orientalised Hellenism versus Hellenised Orient: Reversing the Perspective on Gandharan Art. *Ancient Civilizations from Scythia to Siberia*, 18, 111–141. DOI: <https://doi.org/10.1163/157005712X638663>
- Foucher, A.** (1905–1951). *L'art gréco-bouddhique du Gandhāra; étude sur les origines de l'influence classique dans l'art bouddhique de l'Inde et de l'Extrême-Orient*. (3 vols.). Paris: Imprimerie Nationale.
- Jongeward, D., Lenz, T., Neelis, J., & Pons, J.** (Forthcoming). *Buddhist Rebirth Narratives in Literary and Visual Cultures of Gandhāra*. Seattle: University of Washington Press.
- Kahle, P., Colutto, S., Hackl, G., & Mühlberger, G.** (2017). Transkribus—A Service Platform for Transcription, Recognition and Retrieval of Historical Documents. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* (pp. 19–24). DOI: <https://doi.org/10.1109/ICDAR.2017.307>
- Miles, A., & Bechhofer, S.** (2009). *SKOS simple knowledge organization system reference* [W3C recommendation]. W3C. <https://www.w3.org/TR/skos-reference/>
- Pons, J.** (2011). *Inventaire et Étude Systématiques Des Sites et Des Sculptures Bouddhiques Du Gandhāra: Ateliers, Centres de Productions*. Unpublished PhD thesis.
- Pons, J.** (2017). Archaeology in Gandhāra: A Review of Research at the Crossroads of Disciplines. In A. Lichtenberger & R. Raja (Eds.), *The Diversity of Classical Archaeology (Studies in Classical Archaeology 1)* (pp. 199–219). Turnhout: Brepols Publishers.
- Salomon, R.** (2018). *The Buddhist Literature of Ancient Gandhāra. An Introduction with Selected Translations*. Summerville: Wisdom.
- Strong, J.** (2001). *The Buddha: A short biography*. Oxford: Oneworld.
- The J. Paul Getty Trust.** (2014). *Getty Vocabularies: LOD. AAT Semantic Representation*. https://www.getty.edu/research/tools/vocabularies/lod/aat_semantic_representation.pdf
- Taddei, M.** (1980). “Buddha e Apollo”. In A. Semino (Ed.), *Le grandi avventure dell'archeologia* (pp. 1943–1964). Rome: Armando Curcio Editore.
- Zwalf, W.** (1996). *A Catalogue of Gandhara Sculptures in the British Museum* (2 vols.). London: British Museum.

TO CITE THIS ARTICLE:

Autiero, S., Elwert, F., Moscatelli, C., & Pons, J. (2023). The Seven Steps: Building the DiGA Thesaurus. *Journal of Open Humanities Data*, 9: 11, pp. 1–14. DOI: <https://doi.org/10.5334/johd.111>

Submitted: 12 May 2023

Accepted: 16 June 2023

Published: 31 July 2023

COPYRIGHT:

© 2023 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.