Annamaria De Santis and Irene Rossi (Eds.)
Crossing Experiences in Digital Epigraphy. From Practice to Discipline

Annamaria De Santis and Irene Rossi (Eds.)

Crossing Experiences in Digital Epigraphy

From Practice to Discipline

Managing Editor: Katarzyna Michalak

Associate Editors: Francesca Corazza
and Łukasz Połczyński

Language Editor: Rebecca Crozier

DE GRUYTER

ISBN 978-3-11-060719-2
e-ISBN 978-3-11-060720-8

Published by De Gruyter Poland Ltd, Warsaw/Berlin
Part of Walter de Gruyter GmbH, Berlin/Boston
The book is published with open access at www.degruyter.com.

Library of Congress Cataloging-in-Publication Data
A CIP catalogue record for this book has been applied for at the Library of Congress.

Managing Editor: Katarzyna Michalak
Associate Editors: Francesca Corazza and Łukasz Połczyński
Language Editor: Rebecca Crozier

www.degruyter.com
Cover illustration: Łukasz Połczyński; Ancient South Arabian inscription (Ṣanʿāʾ, Military Museum, MṢM 149)

Contents

Introduction

Epigraphy is a multifaceted discipline. Even more than in manuscript studies or papyrology, a researcher approaching an epigraph should be competent with philology, linguistics, archaeology, history of art, not to speak of history tout-court, being inscriptions studied first of all as primary historical sources. The peculiar nature of the epigraphic document – both textual and physical – has put the reflection on digitization of epigraphs at the crossroads of the discussions and advancements in digital humanities and digital heritage, in addition to computational linguistics.

The digitization of the epigraphic heritage is at an advanced stage. A significant number of projects digitizing inscriptions, of both small and big corpora, with different objectives are either under development, or have been recently completed. Many papers have been written, and several proceedings of meetings and conferences dedicated to this topic have been published.

However, digital epigraphy is not yet considered a proper discipline. Digital epigraphers have acquired their skills in digitization methods and techniques informally, "in the field", through a progressive refinement of those established in the digital humanities. Scholars interested in digital epigraphy are creating more or less formal networks in order to exchange ideas and suggestions, even in very different historical and geographical domains. Nevertheless, there are still no regular occasions to meet and discuss.

Moreover, this large and across-the-board community does not recognize itself in specific journals. They continue to communicate the results of their scientific and technical activities in journals dealing with traditional epigraphy, or, at best, digital humanities in general.

This book is precisely intended to stimulate debate among those practicing digital epigraphy, by recording the methodological issues they have addressed while carrying out specific projects, the solutions they have applied and the criteria that have led to their choices.

In particular, whereas a consistent number of digital initiatives in the domain of Classical epigraphy have been well represented in the proceedings of conferences organized within the frame of the project EAGLE,[1] other domains – and that of Semitic epigraphy *in primis* – are in a quite different situation. Barriers due to the extreme wealth, and also diversity, of writing systems and languages, and to cultural and historical fragmentation, make confrontation and cooperation difficult.

For this reason, the projects represented in the nineteen contributions collected in this book are intentionally diverse in geographic and chronological context, for script and language, and typology of digital output.

1 See further on in the volume (in particular the contributions by Liuzzo) for detailed bibliography.

The Experience of DASI Project

The idea of a volume collecting different experiences of projects on digital epigraphy has arisen within the frame of DASI – *Digital Archive for the Study of pre-Islamic Arabian Inscriptions*, an ERC – Advanced Grant funded project led by Prof. Alessandra Avanzini at the University of Pisa, aimed at gathering, in an open-access archive, the curated edition of the epigraphic corpora of pre-Islamic Arabia. These consist of thousands of Ancient South Arabian, Ancient North Arabian and Aramaic inscriptions produced since the beginning of the first millennium BCE until the advent of Islam. The study of these inscriptions is essential in order to fill a significant gap in research on the ancient and late antique Near East.

During the five years of the project, a team (consisting of epigraphers, archaeologists, art-historians, digital humanists and IT specialists) worked together, facing methodological and technological challenges while building upon previous experiences of digitization of inscriptional corpora in Semitic languages and alphabetic scripts.

Basic, common issues concerned the modelling of data in order to best describe the complex nature of the epigraphic source, and the encoding of text for its critical edition. Fundamental issues such as those of compliance to standards, interoperability and data openness were tackled. Moreover, specific methodological and technical challenges were faced when approaching the study of under-resourced languages, such as those of pre-Islamic Arabia, which are documented only by epigraphic sources. Specific, lexicographic tools were designed to enhance the description of the language and thereby reach a better comprehension of the messages conveyed by the inscriptions – ultimately leading to the best possible understanding and dissemination of the history and culture of the peoples inhabiting Arabia in pre-Islamic times.

The DASI project has attempted to make the tradition of studies related to pre-Islamic Arabia less "marginal" than before, making the edition of about 10,000 inscriptions originating from ancient Arabia openly available. It has tried to provide useful tools and suggest new approaches to the study of this rich cultural heritage, and to foster reasoning on best practice by taking account of domain-specific questions. This has led to a constant search for confrontation with other digital epigraphy projects.

This volume, conceived during the post-grant phase of the project, continues the mentioned practice of confrontation, wishing to raise new questions and open further, unexpected research perspectives.

Concept and Content of the Volume

With this vision in mind, this book gives voice to those who have conceived and carried out diverse projects, ranging: from antiquity to medieval and modern times; from alphabetic to logographic writing systems; from Indo-European to Chamito-Semitic to Ancient American languages; from specific databases and lexica, to aggregators, infrastructures and gazetteers.

Hereafter, summaries of the main characteristics of each project and the topics of the related papers are provided in order to facilitate the readers' orientation.

Chapter 1, by Avanzini, De Santis and Rossi, describes the project DASI – *Digital Archive for the Study of pre-Islamic Arabian Inscriptions*, focusing on the main digital epigraphy themes discussed throughout this volume: text encoding and data modelling, interoperability, and lexicography.

The project RuneS – *Runic writing in the Germanic languages* (Chapter 2) collects texts in different Germanic languages and using different Runic writing systems. This comparative approach to the study of the script has led, as explained in the contribution by Zimmermann, Kazzazi and Bahr, to transcend the existent descriptive systems and enhance the visual documentation of inscriptions, through the tagging of images.

Similarly, *Hesperia – Banco de datos de lenguas paleohispánicas* gathers inscriptions and coins in the different Palaeohispanic languages, written in multiple writing systems. The solutions adopted to register and make searchable both script variants and the different transliterations used in the study tradition, are described by Estarán, Beltrán, Orduña and Gorrochategui in Chapter 3.

The two projects *Sinlequiunnini* (Di Filippo) and *Text Database and Dictionary of Classic Mayan* (Prager, Grube, Brodhun, Diederichs, Diehr, Gronemeyer and Wagner) propose different solutions in the textual data modelling in relation to logo-syllabic writing systems, in particular dealing with languages whose interpretation is highly context-driven, in the first case (Chapter 4), and with a still partially deciphered script, in the second one (Chapter 5).

The *Beta Maṣāḥəft* project (Chapter 6) deals with Ethiopian and Eritrean inscriptions and manuscripts. Bausi and Liuzzo address the issue of encoding in XML the relation among multiple copies of the same epigraphic text in a multilingual context, and of annotating their different scripts.

The CIP – *Corpus Inscriptionum Phoenicarum necnon Poenicarum* (Chapter 7) is the first attempt at carrying out a census of the Phoenician and Punic inscriptions spread in a very wide territory, from the Eastern to the Western Mediterranean. The contribution by Xella and Zamora provides an overview of the criteria they have followed to create a complete edition of the only direct textual sources for the reconstruction of the history and culture of this civilization, in the current absence of any attestation of literary texts.

The OCIANA – *Online Corpus of the Inscriptions of Ancient North Arabia* project (Burt, al-Jallad and Macdonald) is a database mainly designed to catalogue graffiti.

Their curated editions, including transcriptions, transliterations in Latin characters and translations, include encoding with particular attention to grammatical analysis and onomastics (Chapter 8).

As the mentioned projects show, the digitization of the overall epigraphic heritage is often aimed at supporting linguistic study. The *Sabaic Dictionary Online* aims at cataloguing all extant lexical material of one of the Ancient South Arabian languages (Chapter 9). Multhoff provides a sound explanation of the methodological issues concerning the annotation of morphological analysis: treatment of ambiguous forms, homographs, heterographs with identical meaning, variant readings, incorrect forms.

The lemmatizer for the Ancient South Arabian languages, KALAM, performs the automatic detection of morphological attributes (Chapter 10). Ruzicka describes its principles and functioning. The contribution must be considered within the frame of the application of NLP techniques to ancient, under-resourced languages.

The OIMEA – *Official Inscriptions of the Middle East in Antiquity* project (Novotny and Radner) edits all the official inscriptions of ancient Middle Eastern polities in cuneiform script. Texts are geo-referenced and fully lemmatized: lexical and grammatical tagging is carried out in order to create glossaries and allow search of text and translation. Historical research is enhanced by the creation of a map-based interface to access geographical information mentioned in cuneiform sources (Chapter 11).

The project *Karnak* (Biston-Moulin and Thiers) focuses on the epigraphs located *in situ* in the ancient Egyptian temples of Karnak. Therefore particular attention is devoted to the preservation of the relation between the inscriptions and their architectural position. An extensive photographic coverage provides high-resolution orthophotographs flanking the transliterations of hieroglyphic, hieratic and demotic texts. These are the basis for a digital lexicon of the languages documented in the temples (Chapter 12).

The infrastructure of the HPM – *Hethitologie-Portal Mainz* (Chapter 13) provides maintenance and access to several independent digital resources available on Hittitology studies. Müller and Schwemer recall the history of a long-lasting project; the continuous technical updates that have been necessary over time; the specific policies for the attribution of resources, their versioning and intellectual property.

Other projects cope with the establishment of systems to identify, sort and connect digital resources. The interdependence of geographic and chronological entities and their labelling, and the need for ontologies with the objective of structuring this information is exemplified by the project EDV – *Epigraphic Database Vernacular* (Cannata), which collects the vernacular inscriptions produced in Italy from late Medieval to Early Modern Age (Chapter 14).

The *Trismegistos* project (Depauw) aims at implementing an identification system, which attributes an ID to each known ancient inscription. This is a first step to tackle the issue of disambiguating and connecting several editions for the same inscriptions in a LOD environment (Chapter 15).

The objective of the project *PeriodO* (Rabinowitz, Shaw and Golden) is the creation of a Linked Data gazetteer of structured period definitions, which provides links between time periods and geographic and cultural contexts, and translation between absolute dates and relative chronologies. Once applied to digital epigraphy, it will foster interoperability of epigraphic collections and their connection with archaeological datasets (Chapter 16).

Interoperability is fully achieved by the aggregator EAGLE, which collects Greek and Latin epigraphs from many different repositories and makes them available to Europeana. The contribution by Liuzzo focuses on the challenges faced, during and after the end of the project, from the up-conversion to the EAGLE schema of the epigraphic records to the harmonization of the terminologies involved (Chapter 17).

Finally, the EPIDAT – *Database of Jewish Epigraphy* project (Kollatz; Chapter 18), which provides its records to national and European aggregators not specifically focusing on digital epigraphy, and the I.Sicily – *Inscriptions of Sicily* project (Prag and Chartrand; Chapter 19), which, in addition to a consistent amount of previously undigitized epigraphs, provides original editions based on the principles of reuse, linked data and collaboration, demonstrate the potential of records encoded according to the best practice shared by the scientific community.

The volume is provided with an index, listing terms grouped by: Ancient and Modern Regions and States; Languages and Scripts; Concepts of the epigraphic discipline and related digital practice. Finally, two appendices complement the volume. Appendix A presents an annotated webliography of selected online electronic resources cited in the volume, described according to the Dublin Core Metadata Element Set (Version 1.1). Appendix B is intended for disambiguation and definition of selected concepts from the Index of Concepts, by mapping them to the Library of Congress Subject Headings and the Getty Art and Architecture Thesaurus.

Reading Path

The deliberate heterogeneity of subjects, focuses and approaches to digital epigraphy represented in this volume, allows a non-sequential fruition of the contributions. However, they are grouped into two main subject areas. These areas, which have been part of the research of DASI itself, enclose, in our opinion, the main issues that digital epigraphy should address in developing a methodology able to provide the validity criteria proper to a discipline.

1. The first part of the volume is focused on data modelling and encoding, which deeply influence the possibility to perform searches on texts including *lacunae* and variants.
   - Various scripts, belonging to different writing systems and often not completely deciphered, pose fundamental issues in relation to data modelling and/or encoding, given the high uncertainty in the attribution of

phonetic, morphological and semantic values to graphemes and sequences of graphemes.
- Data modelling and encoding are also influenced by the will of creating proper critical editions of epigraphs and the specific functionalities required to meet their criteria.
- Moreover, different languages, often extinct and not completely understood in their morphology and lexicon, need to be studied from the linguistic point of view, before historical, cultural, sociological and much more interpretation can be derived. Lexica and tools for morphological analysis, specifically developed on the basis of the epigraphic collections digitized, and coping with fragmentarily attested languages, are therefore described.

2. Interoperability and aggregation are fundamental to relate data that would otherwise remain separate, in contrast to the reality they refer to. This second part of the volume is dedicated to the initiatives aimed at fostering aggregation, dissemination and reuse of epigraphic materials. It includes:
   - the experiences which point out the need, and tools, for interoperability
   - portals providing "annotated" access to several digitization projects, and proper aggregators
   - and projects which, thanks to interoperability, are clear examples of successful dissemination of inscriptions digitized in different projects.

Although the contributions allow multiple keys to interpretation, and the editors encourage a "personal" fruition, this ordering of the papers aims at suggesting a reading path. This path follows the red thread of the dialectical relationship between the need to represent in the digital environment the features of peculiar epigraphic materials in the most effective way, and the need for strategies to share, disseminate, and make data reusable. In other words, the relationship between the compliance with the theoretic tools and the methodologies developed by each different tradition of studies, and, on the other side, the necessity of adopting a common framework in order to produce commensurable and shareable results in digital epigraphy.

In sum, by crossing a wide, even though not exhaustive, range of experiences, this volume intends to point out the methodological issues which are specific to the application of information technologies to epigraphy. It was not conceived to be a prescriptive work; it does not provide answers, but focuses on problems. Eventually, it aims at stimulating interest and discussion around the challenges that the use of IT has been imposing on epigraphy and on how the digital approach is reshaping the very discipline.

Acknowledgements

The dedication we have put in the preparation of this volume is our personal homage to the Director of DASI project, Prof. Alessandra Avanzini. Under her guidance, DASI has been for us a time of absorbing study, passionate debate, curious experimentation, continuous rethinking and multidisciplinary encounters with the many people of the DASI research groups of the Scuola Normale Superiore and the University of Pisa, who have shared with us daily enthusiasm and hard work.

We take this opportunity to thank the contributors to this volume, who have dedicated, in spite of their many commitments, time and energies to reflect on common issues and challenges.

Our thanks go also to the editors of De Gruyter, and in particular to Katarzyna Michalak, for their constant assistance.

The FP7 post-grant Open Access Pilot project has provided financial support for the publication of the present volume.

Annamaria De Santis and Irene Rossi

Alessandra Avanzini, Annamaria De Santis and Irene Rossi

1 Encoding, Interoperability, Lexicography: Digital Epigraphy Through the Lens of DASI Experience

Abstract: This paper describes the main challenges faced and the solutions adopted in the frame of the project DASI – *Digital archive for the study of pre-Islamic Arabian inscriptions*. In particular, it discusses the methodological and technological issues that emerged during the conversion from the CSAI – *Corpus of South Arabian inscriptions* project (a domain-specific, text-based, digital edition conceived at the end of 1990s) to the wider DASI archive for the study of inscriptions in different languages and scripts of ancient Arabia. The paper devotes special attention to: the modelling of data and encoding (XML annotation vs database approach; the conceptual model for the valorisation of the material aspect of the epigraph; the textual encoding for critical editions); interoperability (pros and cons of compliance to standards; harmonization of metadata; openness; semantic interoperability); lexicography (tools for under-resourced languages; translations), with a view to possibly fostering reasoning on best practices in the community of digital epigraphers beyond each specific cultural/linguistic domain.

Keywords: data modelling, text encoding, interoperability, lexicography, pre-Islamic Arabia

1.1 Digitizing the Epigraphic Heritage of Ancient Arabia: From CSAI to DASI

From the beginning of the first millennium BCE, in the region corresponding roughly to modern Yemen and neighbouring areas in Oman and Saudi Arabia – the so-called *Arabia Felix* of the classical sources – the Ancient South Arabian civilization flourished. During a long history of more than 1,500 years, the Ancient South Arabian four main kingdoms of Maʿīn, Saba, Qataban and Ḥaḍramawt produced a written documentation currently consisting of around 15,000 inscriptions, which constitute the direct textual source for the knowledge of the Ancient South Arabian civilization, as no literary texts have been discovered yet (Avanzini, 2016).

Recognising the need for a systematic collection of this epigraphic heritage, in 1999 Prof. Alessandra Avanzini at the University of Pisa undertook the project of an

Alessandra Avanzini, Annamaria De Santis, Università di Pisa
Irene Rossi, Consiglio Nazionale delle Ricerche, Roma

online *Corpus of Ancient South Arabian Inscriptions* – CSAI (Avanzini, Lombardini, & Mazzini, 2000). The choice of producing an online curated textual corpus – even before considering its paper edition (Avanzini, 2004) – was determined by several advantages that apply to any cultural domain of study, but that are especially indispensable for those "young" disciplines, whose progress determines a constant re-definition of previous theories. Those advantages are: the updatability and expandability of the collection, the potential improvement of the edition of the sources and of the consultation tools, including full-text retrieval tools, the immediate accessibility of the material – published in scattered, often inaccessible publications, or coming to light from excavations at a fast pace – and its potentially infinite dissemination.

The CSAI archive, realized with the technical support of the Scuola Normale Superiore di Pisa, went online in 2001. Its starting bulk was comprised of some 1,300 texts of the *Corpus of Qatabanic Inscriptions*. The archive content was continuously updated for a decade, so to comprise the whole collection of Qatabanic, Minaic and Ḥaḍramitic inscriptions, plus a number of Sabaic texts – Sabaic being the most consistent South Arabian epigraphic corpus (Figure 1.1).

Figure 1.1: CSAI homepage (2010)

Related, funded projects aimed at the cataloguing of not just the Ancient South Arabian, but also the Nabataean and Phoenician collections of inscriptions and

artefacts preserved in museums worldwide,[1] allowed the content of the archive to be enriched. These projects also enhanced a continuous methodological reflection and technical elaboration, allowing a definition of best practice and development of tools for the study of a peculiar documentation, whose state of research is still "fluid".

It is precisely from this kind of experience, that some ten years later a wider project, the Digital Archive for the study of pre-Islamic Arabian inscriptions (DASI), was conceived and funded with an ERC Advanced Grant awarded to Prof. Avanzini. The objective was to enhance knowledge of the history, language and culture of the whole of ancient Arabia by studying its textual heritage; a heritage that is composed of tens of thousands of inscriptions in the Ancient North Arabian, Ancient South Arabian and Aramaic languages and scripts.

Both the digitization tool and archive's public website of the CSAI were re-designed, in order to conform to the new research objectives of the DASI project and to the advancements in digital humanities that had occurred during the last decade (cf. in general Schreibman, Siemens, & Unsworth, 2004; Babeu, 2011). The process of re-engineering a system which already contained a large amount of data (around 6,000 inscriptions, with encoded text, metadata, translations, bibliographical references and visual documentation) and the migration of structured data, brought to light a series of methodological and technical issues. Only part of them could be satisfactorily faced.

In the present paper, the main challenges we encountered, the proposed solutions, the still open questions and the prospects we envisage for the future of digital epigraphy – starting from our experience within the DASI project – will be discussed, dealing with three core themes: data modelling and text encoding, harmonization and interoperability, and lexicography.

1.2 Data Modelling and Textual Encoding

1.2.1 The Data Model: XML vs Database

During the 1990s, textual encoding was successfully experimented with literary sources, and became the standard approach for projects interested in digitizing and annotating texts. The IT system of the CSAI was developed by the "Centro di Ricerche Informatiche per i Beni Culturali" (CRIBeCu) of the Scuola Normale Superiore of Pisa, which had acquired specific know-how in the field of text encoding and had

1 MENCAWAR – Mediterranean Network for Cataloguing and Web Fruition of Ancient Artworks and Inscriptions [http://arabiantica.humnet.unipi.it/index.php?id=mancawar]; CASIS – Cataloguing and Fruition of South Arabian Inscriptions through an Informatic Support [http://arabiantica.humnet.unipi.it/index.php?id=casis].

developed TReSy (acronym for Text Retrieval System). This was one of the early full-text SGML-XML search engines, able to perform accurate queries on the context (Lini et al., 2004). Metadata and texts of the CSAI inscriptions were recorded in SGML, and later XML files, according to a schema specifically created for CSAI. Indeed, best practice and standards, such as those of TEI and EpiDoC, were not yet widespread, especially in Europe.

This kind of approach, centred on the manipulation of the text, suffered from a range of shortcomings in the description of the text-bearing object and in the management of complementary resources such as bibliographical records and visual documentation. Moreover, the system forced users to handle the XML, often discouraging potential encoders, and did not allow the control of the workflow and the real-time updating of data.

To overcome these limitations, a new system was designed for the DASI project by the staff of the Scuola Normale Superiore, consisting of a web based, relational database enabling a controlled and swift workflow by different levels of authorization for each curatorial role, and uniformity of data by an extensive use of lists of controlled terms, editable by authorized users.

An XML editing module for the textual transcription and encoding of the pre-Islamic Arabian inscriptions was integrated into the database. This is provided with a set of buttons to enter the annotation of all, and only the phenomena considered within the project, ensuring easiness and uniformity of mark-up. The validity and well-formedness of the documents against the schema are granted by preventing elements being entered in incorrect positions, and by managing overlapping of tags through a system of identifiers and couplings between the fragments of the broken elements. The entire content of the database – text encoded and metadata – is then extracted in XML by a web service, in order to construct the dynamic sections of the front-end.[2]

In the context of a "niche" discipline, the design of easy-to-use tools such as the DASI XML editor, as well as the entire data entry system (Figure 1.2), was an effective step towards a wider involvement of scholars in the digitization and curatorial tasks. Moreover, DASI system has proved to be a performing didactic tool in the teaching of epigraphic disciplines and Semitic languages. The virtual keyboard with diacritic characters helps in the transliteration, and the scientific terminology displayed on menus and buttons for textual mark-up suggests coherent definitions to be used to. The process of encoding develops the students' familiarity with methods and tools of philological and linguistic analysis.[3]

2 [http://dasi.cnr.it]. DASI IT system is currently maintained by the CNR Reti e Sistemi Informativi, with the scientific supervision of the CNR Dipartimento Scienze Umane e Sociali, Patrimonio Culturale.
3 Cf. Bodard & Stoyanova, 2016 for similar experiences in the domain of Classic epigraphy.

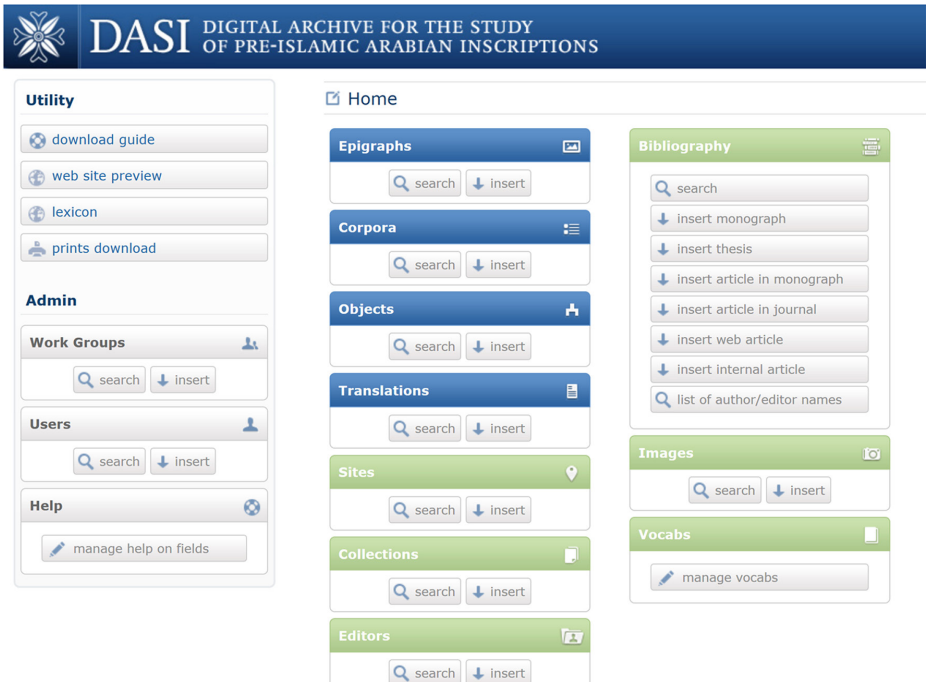

Figure 1.2: DASI data entry interface

1.2.2 The Conceptual Model: Text vs Object

Given the obvious focus of the DASI project on the inscriptional text, as any other epigraphic project, the Epigraph entity (see below) is the most articulated one in the conceptual model of the database. Besides the XML editor for textual transcription and annotation, it contains the metadata of the text (on linguistic features, writing, chronology, genre, notes of *apparatus criticus*, general and cultural notes).

Metadata and text of the inscription's translation(s) are recorded in the related Translation entity. Additional entities complete the description with geographic information (Site), visual documentation (Image), references to the history of studies (Bibliography) and indications of curatorial responsibilities within the DASI project (Editor).

The core issue in the conception of the DASI model was the need to account for and valorise the material aspect of the epigraphic document. As stated above, in the traditional encoding approach this proved to be under-represented in comparison to the textual aspect, to such an extent that information on the supports of the inscriptions was encoded as metadata of the text.

Therefore, the innovation in the DASI approach, compared to the CSAI, is the separation of the information concerning the text from that concerning its physical support in two different but related entities. The recording of the archaeological and historic-artistic information on text supports in a dedicated Object entity, allows the additional problem of the multiplication of object records in the case of objects bearing multiple, self-contained texts to be overcome, and the one-to-one relation between the object in the database and the real object to be maintained. Moreover, the autonomy to the Object entity allows to record uninscribed objects, with the additional outcome of enhancing the study of the history of art of pre-Islamic Arabia and valorising specific museums' collections of objects in the DASI archive.

The DASI website reflects the text-object distinction, via the two main indexes of *corpora* and *collections* that group texts and supports on the basis of their linguistic attribution or current deposit respectively. This has proved extremely important for the preservation and valorisation of the Yemeni cultural heritage, as some of the museums' collections catalogued in DASI have undergone serious damage or pillage, or were entirely destroyed during the ongoing war.[4] Securing the existence of digital copies of objects at risk from environmental or human factors is today of primary importance. We believe that their description as well as their visual documentation – and the open access and re-usability of both – should be among the major concerns of projects involved in the digitization of cultural heritage, for preservation purposes.

The distinction proposed in the DASI model between texts and supports, though suitable from the conceptual and practical points of view, has its limits due to the strict relation between them (e.g. the spatial relations among components of the text, the distribution of text on the support, the relation of the texts with the iconographic elements and decoration), and with the communicative context. The case of the monograms is emblematic. The monogram is not an abbreviation inside the text, but a combination of signs decorating an object (Figure 1.3). The same monogram may occur engraved next to a text or even without a linear inscription. In many cases, the name the monogram refers to is unknown, because some letters can be omitted or incorporated into the shape of other letters, there being no way to reconstruct their correct order in these symbolic representations. Therefore, are monograms inscriptions, or rather decorations, of objects? Should they be encoded in the Epigraph or described in the Object?

A further example of the relation between texts and supports is the mention within the epigraphic text of the type of object on which it is inscribed. The correspondence between the term and its material signifier is extremely relevant for the improvement of the knowledge of both the pre-Islamic Arabian languages and the material culture. However, the data model adopted does not allow for a direct correlation between

4 [http://en.unesco.org/galleries/heritage-risk-yemen].

them, nor between parts of the text and their visual documentation, which may be improved through the tagging of images.

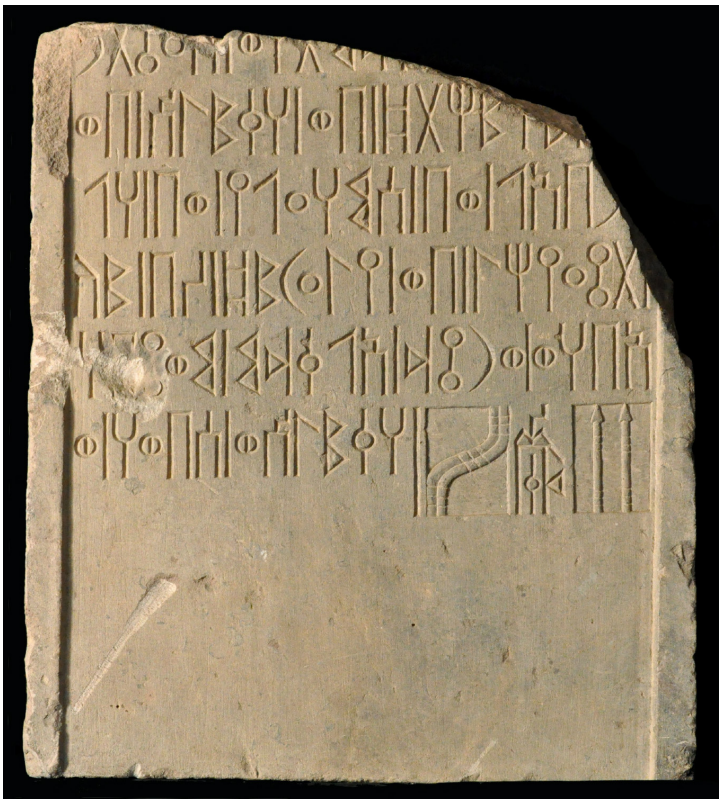

Figure 1.3: Early Sabaic boustrophedon inscription with monogram and symbols (MṢM 149)

1.2.3 Encoding for Curated Digital Editions: In-Line vs External Apparatus Criticus

The XML editor integrated in the DASI data entry system (Figure 1.4) allows encoding of texts in compliance with the EpiDoc subset of the TEI[5] standard (Elliott et al., 2007–2016). The annotated phenomena are linguistic (onomastic, grammatical), philological (lacunae, restorations, corrections, etc.), descriptive of the relation between text and support (line breaks, text turning around the object) or of the internal structure of the text (genealogies, eponyms), etc. The critical notes are collected in a separate

5 Text Encoding Initiative [http://www.tei-c.org/].

section and refer to the concerned text by the indication of the corresponding line of the transliteration – a traditional approach of managing the *apparatus criticus* that has been inherited from the project CSAI.

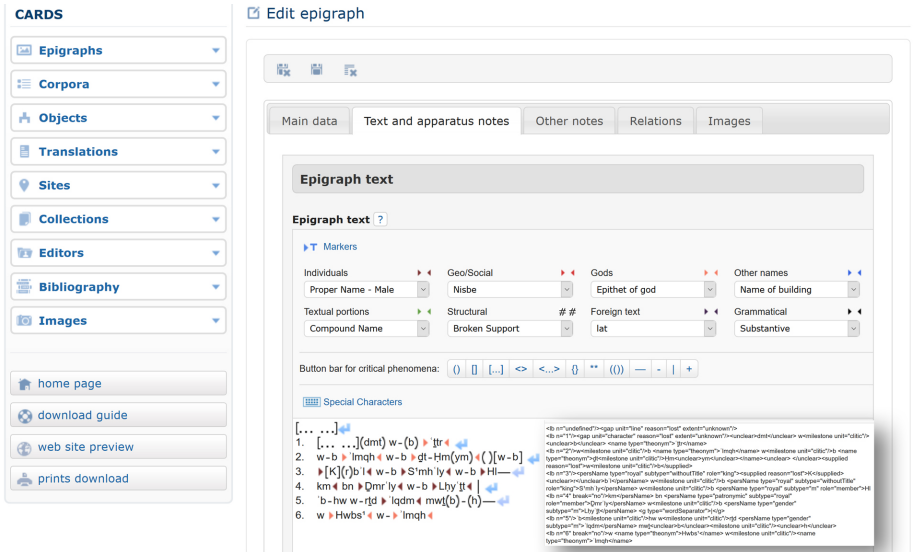

Figure 1.4: DASI XML editor

The solution many projects have adopted in order to valorize the apparatus notes is the encoding of the text contained in them, and its referencing to the corresponding section of the epigraph transcription. The alternative solution is the insertion of the *apparatus criticus* in-line, directly in the transcription's annotation. This is particularly interesting, as it allows retrieval, through a textual search, of all the possible readings/interpretations of a textual passage, or the renderings of the texts suggested by different editors. Indeed, the *<app>* with *<lem>* and *<rdg>* elements have been used in the DASI XML editor to encode variants of uncertain readings or of restorations, or of linguistic (mainly onomastic) interpretations, when none of them could be discarded.[6] As it is apparent, the main concern in the DASI in-line encoding of variants is not so much to retrieve single variants of words, as to retrieve them

6 These elements were created in the TEI to encode the variants occurring in a work's multiple witnesses, as in the case of manuscripts. However, their semantic value can be applied to encode information on different critical editions of one epigraph, because the strong emphasis on the physical nature of an epigraph leads to consider each inscription as a unique specimen. This solution is presently suggested also in [http://www.stoa.org/epidoc/gl/latest/supp-app-inline.html]. EpiDoc guidelines in general are available at [http://www.stoa.org/epidoc/gl/latest/].

within a specific context, consisting in the portion of text preceding and following the text characterized by variants.

The tool for combined searches on text and extra-textual data provided in the DASI website allows queries on words or word patterns within a phrase, with the possibility of setting the maximum number of words that can intervene between the first and the last words searched for (Avanzini, Prioletta, & Rossi, 2014). The search can be restricted to lexical or onomastic results – even within a particular onomastic category. The adoption of an in-line approach in recording the *apparatus criticus* of the inscriptions would exponentially augment the potential of such a search tool.

However, to encode the *apparatus criticus* in-line at such a level of detail as to provide an "encoded history of study" of an inscription (i.e. rendering on one single file the interventions applied by different scholars in their own edition of the text) is a very long and complex task. Moreover, it entails the risk of over-tagging the transliteration of the text by applying too many "layers" on it. On the other hand, providing several files for the different editions of the same inscription, to be then grouped within aggregators, is a viable solution, but it limits the potentialities offered by a digital edition.

1.3 Interoperability

1.3.1 Text Encoding and Representation: Standards vs Specificities

All of the scripts used to write down the inscriptions considered within the DASI project (Ancient South Arabian, Ancient North Arabian and Aramaic varieties) are alphabetic. In Southern Arabia a geometric, monumental writing is evidenced since the 9th–8th century BCE by the "public" inscriptions: each letter is graphically separated from the adjacent ones and the division between the orthographic units (which, as typical of the Semitic languages, can be composed by a main word plus affixes for clitic pronouns, conjunctions, particles) is marked by a vertical trait. A "cursive" writing was also in use to record private, movable or archival texts on wooden sticks (contracts, lists of goods, correspondence, school exercises, etc.). As the majority of Semitic scripts, the ductus of writing is normally right-to-left, although in ancient South Arabia there are a considerable number of boustrophedon inscriptions as well. The Ancient North Arabian texts – except for a few hundred "monumental" texts from major settlements – consist mainly of graffiti left by nomadic people on desert rocks, and their direction of script is much more varied, sometimes even circular.

The inscriptional text is entered in the DASI XML module in Latin transliteration, using the UTF-8 set of the Unicode standard. The transliteration of Semitic phonemes in Latin characters implies the addition of diacritical marks (like underdots) to the letters and therefore discourages the representation of editorial phenomena according to the Leiden conventions: the latter, elaborated in the frame of Classical philology

and recommended by EpiDoc, visualise the uncertain reading of signs precisely by dots under the letters.

More generally, the DASI project has adhered to the TEI-EpiDoc standard to encode texts, with some limitations imposed by: the need to comply with the specific tradition of studies (choice of phenomena to annotate), the inheritance of the CSAI custom-made encoding schema (already applied to some 6,000 inscriptions) and, related to this, the peculiar interests of the project (like the linguistic, more than prosopographical focus on onomastics). This process of mark-up conversion and the effort towards the alignment to a standard have shown their potentialities in terms of content rethinking and redefinition, and at the same time the need to safeguard as much as possible the specificities proper to each cultural domain and tradition of studies, in order not to lose peculiarities, profoundness and nuances (Avanzini et al., 2016).

1.3.2 Harmonization of Metadata

As explained above, the DASI encoding of texts does not fully comply to the EpiDoc standard's recommendations as regards some transcription phenomena and editorial interventions, and for the encoding of onomastics. However, particular attention has been paid to the harmonization of those metadata elements that entail a reference to structured terminologies. Indeed, the tradition and the state of the art in a discipline exert their influence above all in the classifications that stand at the basis of the knowledge organization systems.

This is exemplified by the lists of controlled terms related to the textual typology and the type of object, which best show the progress in the understanding of the peculiarities of the pre-Islamic Arabian textual tradition and material culture (Avanzini, Prioletta, & Rossi, 2014). The three main typologies of inscriptions – i.e. dedications to the gods, celebrations of construction activities, and legal/administrative regulations – are distinguished by specific formulary patterns (lexical items – in particular the main verb of the inscription, which is the fulcrum of the action described throughout the text – and syntactic features) and very rigid textual structures (the order of the text sections). These were replicated through the centuries, with few areal and chronological variants, and rarely conceded space to the insertion of digressions or to the combination of different textual typologies in the same inscription. The texts encoded in the DASI archive are classified on the basis of those fixed textual models. The comparison with terminologies used in other projects, such as those harmonized in the project EAGLE, has pointed out that some of the entries have exact matches, others are just related to some terms, and the remaining ones have no match at all. This is because different criteria have guided the creation of such classifications and therefore of the vocabularies in use.

Even internally, the DASI project has faced the issue of managing a diversified documentation.[7] The textual encoding accounts for all of the three main language corpora considered within the project (Ancient South Arabian, Ancient North Arabian, and Aramaic), though with obvious compromises as regards specific grammatical features and definitions for each language. It was more difficult to find shared solutions for metadata. For instance, the CSAI project had catalogued and annotated information that was especially relevant to the comprehension of the "monumental" inscriptions (the majority of Ancient South Arabian texts), while most of the Ancient North Arabian inscriptions consist of graffiti. The two categories of texts considerably distance themselves with respect to their scope, audience, authorship, context, etc.; therefore the information that one wants to point out and extract to enhance their study is different. For instance, much attention has to be paid to the artistic description of the support of a monumental inscription, whilst the technique of incision and the relative disposition of texts on a rock are essential information to describe graffiti.

As regards the physical supports of the texts, the specimen that DASI has collected demonstrates its own peculiarities. For instance, stelae make a large part of the artefacts catalogued. Common terminologies, such as the Getty Art & Architecture Thesaurus,[8] include only one term to classify them, but the South Arabian stelae have different, codified morphological and iconographic characteristics that are fundamental (as much as their texts) for the identification of their area of production, dating and function.[9]

Even for those entries that have exact matches, further subcategories may be required to provide specifications useful to scholars interested in a particular domain. In South Arabia, for instance, bases can be found as support to statues, sculptures of heads and stelae. Their morphological and functional – i.e. communicative, not only material – features, as well as the geographical and chronological distribution, may vary considerably. Is it possible to increase the granularity of the shared terminologies without reproducing the domain-specific typologies of the classes of materials? For instance, let us consider the bases of Ancient South Arabian statuettes, which have been found in temples for propitiatory and votive aims. We would consider it inappropriate to map the South Arabian bases to a concept having such a domain-

7 When designing the metadata and the tags of the XML editor, the project benefited from the collaboration of colleagues at the CNRS-UMR Orient & Méditerranée as regards the Aramaic corpus, and at the University of Oxford as regards the Ancient North Arabian corpus (see Chapter 8 in this volume).

8 [http://www.getty.edu/research/tools/vocabularies/aat/].

9 For instance, large, rectangular stelae with a decoration of ibexes and *bucrania* framing the text, always bear dedicatory texts and are typical of the Sabaic and Minaic areas, especially in ancient times. Small trapezoidal aniconic stelae whose base is inserted on an inscribed plinth, as well as rectangular, beautifully carved stelae with the representation of the deceased's bust and his/her name inscribed below the figure, usually bear Qatabanian funerary texts.

driven definition as the bases of statues in the Classic world, which are placed in the public, civic space with honorary function.

1.3.3 Openness and Semantic Interoperability

In relation to the public funding of the project and the policy adopted by the EC on Open Access to publications and research data, the DASI project has made available the entire archive in open-access modality. The DASI repository allows service providers to harvest its records through the OAI-PMH protocol (Avanzini et al., 2015).[10] As the archive is not an aggregator in the strict sense, the DASI project has developed a general data model able to convey an accurate description of the material support, the historical and geographic context, and the textual content of the pre-Islamic inscriptions of the Arabian Peninsula, but not a proper schema. Therefore, the key point has been mapping the DASI data model to the DC elements set, as required by the OAI-PMH protocol, and to the EDM in order to expose records to the Europeana aggregation service, in addition to the mentioned EpiDoc subset.

A further step to achieve semantic interoperability,[11] in addition to interoperability at the repository level and at the record level, is related to the names of individuals and places. The DASI encoding of onomastic phenomena is detailed and articulate. However, for the time being, it has had a linguistic objective rather than a prosopographical one. The royal onomastics is easily recognizable and extremely repetitive, as it was probably taken on with the investiture. Genealogies of kings are therefore rather evanescent, so much to suggest that the institution represented was more important than the individual king, at least until the last centuries BCE (Avanzini, 2016, pp. 53–57). Then, it is difficult and highly hypothetical to identify a single person, place him/her over time, and relate with certain attestations. Nevertheless, it would be worth seeking to do this for the main historical figures and for some periods, for instance when inscriptions begin to be dated and therefore the identification of individuals is less tentative.

Similar considerations could be made about place names. The DASI onomastic lists include about 3,600 names of geographical, social and political entities that have been tagged in the epigraphs: elements of the natural and the human landscape, entire settlements and single artifacts (buildings and monuments), political and social entities (states, tribes, families) which have relations with the territory. Furthermore,

10 DASI repository [http://dasi.cnr.it/de/cgi-bin/dasi-oai-x.pl?verb=Identify].

11 DASI does not apply a frankly semantic approach from the technical point of view, even though the distinction between the physical carrier and the text inscribed in the data model is an implicit result of that way of conceptualizing. However, the mapping of its data to the Europeana Data Model goes in that direction.

archaeological data related to nearly 400 sites, origin or provenance of inscriptions, have been collected: modern and ancient toponyms, including Classical names; country, geographical area and present governorate, coordinates and related accuracy; types of the findings, architectural structures and monuments; chronology; description, history of research; bibliography. Each "Site" record may be linked to the other ones, thus representing the spatial relations among them. A gazetteer is in preparation, which will allow identification and description of all the above-mentioned geographical entities and represent the semantic relations (hierarchy, equivalence and association) among them, in addition to the spatial ones, directly inferred by the primary (epigraphs) or secondary sources (bibliography). This is of particular importance when their actual locations or identification are still unknown.

The difficulties in the historical reconstruction of the pre-Islamic Arabian civilizations are especially apparent at a chronological level, so that the DASI inscriptions are dated to wide periods of three/four centuries. However, as the historical understanding moves forward – and at least for the dated inscriptions since the end of the 1st millennium BCE – an attempt at the semantic interoperability at a chronological level has been envisaged, in connection with the PeriodO project (see Rabinowitz, Shaw, & Golden in this volume).

1.4 Lexicography

1.4.1 Approach to Under-Resourced Languages

Interoperability at a linguistic level across different corpora is a desideratum. The goal of providing useful tools for the research on each of the main corpora that make up the DASI archive (Ancient South Arabian, Ancient North Arabian and Aramaic) has been reached. However, a major issue is still to be approached: whether, to what extent and how to enable combined queries on textual content across documentation in different linguistic families. In fact, not only do these corpora have their own peculiarities (e.g. in terms of language, script, or periodization) that would entail partial or potentially false results, but they also have specific traditions of studies strongly conditioning approaches, methods and definitions. The mapping of grammatical (to a lesser extent) and mainly semantic features of different languages could be one of the ways, though not straightforward and immediate, to facilitate cross-searches on them.[12]

12 During the revision process of the present volume, two books on similar topics have been published. For recent developments and updated bibliography, refer to Juloux, Gansell, & Di Ludovico, 2018, for semantic approach to digital epigraphy of the ancient Near East, and to Cotticelli-Kurras & Giusfredi, 2018, for relations between computational linguistics and digital philology.