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

© 2018 Alessandra Avanzini, Annamaria De Santis and Irene Rossi

This work is licensed under the Creative Commons Attribution 4.0 International license (CC BY 4.0)

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 TRSy (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 (MŞM 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 TEI⁵ 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, Priolella, & 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.

In this sense, the DASI project is implementing lexica of the languages attested in pre-Islamic Southern Arabia, whose current state of knowledge is still very fluid. The development of lexicographic tools starting from the dataset of a digital archive is the ideal situation to advance the research on such fragmentarily attested languages, whose dictionaries and grammatical studies need constant updating due to the growth of sources that are brought to the attention of scholars.

The DASI Lexicon tool has its starting point in the list of words (excluding onomastic items) extracted from the texts encoded (Avanzini et al., 2015). Each of the word forms, corresponding to the items of the words' lists, is retrieved within the contexts of occurrence in the single inscription. One or more rows of translation are in turn linked to each occurrence. Word forms can be assigned, individually or in groups, to a root. While assigning a root, users attribute morphological, part-of-speech (PoS) analysis and translation to the word form. Each word is thus defined, and potential homographic forms are disambiguated (Figure 1.5).

The screenshot displays the DASI Lexicon web interface. At the top, there is a header with the project logo and the text 'LEXICON | CSAI CORPUS OF MINAIC INSCRIPTIONS'. Below the header, there are navigation tabs: 'Admin', 'Edit', 'Preview', and 'Recycler'. The main content area shows a search filter set to 'link status', with a table of 5 different words in a total of 20 occurrences. A modal window titled 'Lexicon - Occurrence editing' is open, showing details for the root 'ghbn'. The modal includes tabs for 'Root form', 'Root meanings', and 'Grammatical analysis'. The 'Root meanings' tab is active, showing a list of meanings for 'ghbn' (bronze) and a section for 'Insert meanings for a new root'. The main table below the modal lists word forms and their occurrences, with links to edit or unlink each entry.

word form	occurrences	lexicon data operations
ghb	☐ DHB (watered valley) / Substantive - Construct - Masculine singular 2 occ. ☐ DHB (bronze) / Substantive - Unclear - Masculine singular 1 occ.	link unlink
ghbn	☐ DHB (bronze) / Substantive - Emphatic/defined by article - Masculine singular 3 occ. ☐ DHB (watered valley) / Substantive - Emphatic/defined by article - Masculine singular 1 occ.	link unlink

Copyright 2013 Università degli studi di Pisa | Scuola Normale Superiore di Pisa | Image Zoom by Cloud Zoom

Project | Contacts | Credits | Terms of use | Privacy

Figure 1.5: DASI Lexicon

This clearly manual approach is mainly related to the constraints of the languages concerned. In fact, the pre-Islamic Arabian languages share with the other, also current, Semitic languages a morphological ambiguity, that is itself a challenge to computational approaches (see Multhoff in this volume). In addition to the typical case of more than one analysis for a given word form, there can be different graphical renderings (spellings) for the same word. Moreover, the high number of *hapax legomena* causes data sparsity. Finally, the small scale of the annotated corpora discourages from effectively driving automatic lexical acquisition.

On the other hand, the remarkable repetitiveness of the texts in relation to the formulaic contexts, suggested not to encode the grammatical and lexicographic information directly on the texts, in order to avoid repeating the editing of all the occurrences. By assigning at one time the same semantic and grammatical analysis to multiple occurrences of a lexical item at a further level, the Lexicon allows completion of the lexicographic work and its potential revision, following the advancements in research, in a reasonable lapse of time.

1.4.2 Translations

While a growing number of digital archives is developing the lexicographic aspect, the majority of them do not include translations, being conceived as traditional collections of primary sources in electronic form.¹³ The project DASI, aiming at the study and the dissemination of the ancient culture of Arabia through the analysis of its epigraphic heritage, conceived its archive as a digital mean for publishing and browsing critical editions of texts, provided with translations, for a better and wider appreciation of their content. More than one translation, often in different languages depending on its bibliographic source, may be linked to one epigraph.

This wealth of data, which takes into account the interpretations by different scholars, however, is not yet encoded. The correspondence between a line of the text and a line of the related translation is a best practice followed by the editor of each item, even though a comprehensible rendering of the concept expressed in the original text often prevents this correspondence, for syntactic reasons. Notwithstanding an effort towards homogenization among the contributions by different editors (at least those directly involved into the project), a strict relation between a word and its translation is also undermined by the semantic differences of words deriving from the same root, or by the semantic nuances one word can acquire on the basis of its context, or more generally by the different morpho-syntactic rules of each language.

Having said that, translations are a key point for searches on texts in different languages. Given the multilingualism of the DASI inscriptional records, translations are going to deserve further methodological reflection and technological effort.

¹³ The awareness of the importance of translations is emerging and several projects, such as AIO “Attic Inscriptions Online” and EAGLE (see Chapter 17 in this volume), have spent much effort on translations to allow a larger public to access epigraphic sources in extinct languages.

1.5 Conclusions and General Remarks

The issues discussed in this paper highlight the efforts that long-lasting digital projects have to make in order to be coherent with their very digital nature, which grants, and at the same requires, constant updating and improving of data, tools and practices.

What epigraphic disciplines, and in general the humanities, are experiencing nowadays, is the contraction both of people engaged in those studies and of funding of projects, creating a vicious circle. Particularly for projects that obtained conspicuous funds, allowing the creation of a large team and undertaking a wide range of initiatives, the abrupt end of short-term grants implies a stalemate. Given the additional efforts required by such a scientific production with respect to more “traditional” outcomes, a major appraisal of the digital products, and specifically of curated editions, in the evaluation process that research and academic staff are subjected to, would stimulate a wider engagement of scholars and early-stage researchers, ensuring sustainability of digital humanities initiatives.

Notwithstanding those apparent difficulties, “young”, “niche” domains of studies, such as the one described in the paper, especially need digital tools, as their state of research entails a continuous production of fresh knowledge and review of theories, and at the same time they are likely to boost the discussion on the perspectives of the digital approach to scientific disciplines – epigraphy in our case – by bringing unexpected issues to the forefront.

Acknowledgements: The research leading to the results presented in this paper has received funding from the European Research Council under the European Union’s Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement n° 269774.

Bibliography

- Avanzini, A. (2004). *Corpus of South Arabian Inscriptions. I-III: Qatabanic, Marginal Qatabanic, Awsanite Inscriptions* (Arabia Antica 2). Pisa: Edizioni Plus-Università di Pisa.
- Avanzini, A. (2016). *By Land and by Sea: A History of South Arabia before Islam recounted from Inscriptions* (Arabia Antica 10). Roma: L’«Erma» di Bretschneider.
- Avanzini, A., De Santis, A., Gallo, M., Marotta, D., & Rossi, I. (2015). Computational Lexicography and Digital Epigraphy: Building digital lexica of fragmentary attested languages in the Project DASI. In G. Guidi, R. Scopigno, J.C. Torres, & H. Graf (Eds.), *2015 Digital Heritage International Congress* (pp. 405–408). New York: IEEE. doi: 10.1109/DigitalHeritage.2015.7419535
- Avanzini, A., De Santis, A., Marotta, D., & Rossi, I. (2014). Between harmonization and peculiarities of scientific domains: Digitizing the epigraphic heritage of pre-Islamic Arabia in the project DASI. In S. Orlandi, R. Santucci, V. Casarosa, & P.M. Liuzzo (Eds.), *Information Technologies for Epigraphy and Cultural Heritage: Proceedings of the First EAGLE International Conference* (Serie antichistica. Collana Convegni 26) (pp. 69–93). Roma: Sapienza Università Editrice. Retrieved from [https://www.eagle-network.eu/wp-content/uploads/2015/01/Paris-Conference-Proceedings.pdf], 2017/11/30.

- Avanzini, A., De Santis, A., Marotta, D., & Rossi, I. (2016). Is still Arabia at the margins of digital epigraphy? Challenges in the digitization of the pre-Islamic inscriptions in the project DASI. In A.E. Felle & A. Rocco (Eds.), *Off the beaten Track: Epigraphy at the Borders. Proceedings of the VI EAGLE International Meeting (24-25 September 2015, Bari, Italy)* (pp. 46–59). Oxford: Archaeopress. Retrieved from [http://www.archaeopress.com/ArchaeopressShop/Public/download.asp?id={E7B2AAC6-9986-4C41-9842-6AA93BE7ACD9}], 2017/11/30.
- Avanzini, A., Lombardini, D., & Mazzini, G. (2000). Corpus of South Arabian Inscriptions. La pubblicazione integrale del corpus sudarabico qatabanico. *Bollettino d'Informazioni del Centro Informatico per i Beni Culturali della Scuola Normale Superiore*, 10.
- Avanzini, A., Priolella, A., & Rossi, I. (2014). The Digital Archive for the Study of Pre-Islamic Arabian Inscriptions: An ERC project. *Proceedings of the Seminar for Arabian Studies*, 44, 15–24.
- Babeu, A. (2011). *“Rome Wasn’t Digitized in a Day”: Building a Cyberinfrastructure for Digital Classicists*. Washington: Council on Library and Information Resources. Retrieved from [https://www.clir.org/wp-content/uploads/sites/6/pub150.pdf], 2017/11/30.
- Bodard, G. & Stoyanova, S. (2016). Epigraphers and Encoders: Strategies for Teaching and Learning Digital Epigraphy. In G. Bodard & M. Romanello (Eds.), *Digital Classics Outside the Echo-Chamber: Teaching, Knowledge Exchange & Public Engagement* (pp. 51–68). London: Ubiquity Press. doi: 10.5334/bat.d
- Cotticelli-Kurras, P. & Giusfredi, F. (Eds.). (2018). *Formal Representation and the Digital Humanities*. Cambridge: Cambridge Scholars Publishing.
- Elliott, T., Bodard, G., Mylonas, E., Stoyanova, S., Tupman, C., Vanderbilt, S. et al. (2007-2016). *EpiDoc Guidelines: Ancient documents in TEI XML (Version 8)*. Retrieved from [http://www.stoa.org/epidoc/gl/latest/], 2017/11/30.
- Bigot Juloux, V., Gansell, A.R., & Di Ludovico, A. (Eds.). (2018). *CyberResearch on the Ancient Near East and Neighboring Regions. Case Studies on Archaeological Data, Objects, Texts, and Digital Archiving*. Leiden: Brill.
- Lini, L., Lombardini, D., Paoli, M., Colazzo, D., & Sartiani, C. (2004). XTReSy: A Text Retrieval System for XML documents. In D. Buzzetti, G. Pancaldi, & H. Short (Eds.), *Augmenting Comprehension: Digital Tools for the History of Ideas*. London: Office for Humanities Communication Publications, King’s College.
- Schreibman, S., Siemens, R., & Unsworth, J. (Eds.). (2004). *A Companion to Digital Humanities*. Oxford: Blackwell. Retrieved from [http://www.digitalhumanities.org/companion/], 2017/11/30.
- TEI Consortium (Eds.). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. TEI Consortium. Retrieved from [http://www.tei-c.org/Guidelines/P5/], 2017/11/30.