

The digital Gazetteer of Ancient Arabia. An example of reuse and exploitation of annotated textual corpora

¹Annamaria De Santis, ²Matteo Gallo, ³Irene Rossi, ⁴Jérémie Schiettecatte

¹Independent researcher, Italy

²Senior developer, Italy

³CNR – Istituto di Scienze del Patrimonio Culturale, Italy

⁴CNRS – UMR8167 Orient & Méditerranée, France

¹annamaria.desantis@unipi.it

²matteogal@gmail.com

³irene.rossi@cnr.it

⁴jeremie.schiettecatte@cnrs.fr

Abstract

Annotated corpora, provided that they adopt international standards and expose data in open format, have many more chances to be easily exploited and reused for different objectives than traditional, analogue corpora. This paper aims at presenting the results of the early adhesion to best practices and principles afterward codified as Open Science and FAIR principles in the frame of projects concerned with digital textual corpora, in a niche area of research such as the pre-Islamic Arabian epigraphy. The case study analysed in this paper is the Digital Archive for the Study of pre-Islamic Arabian inscriptions – DASI, an online annotated corpus of the textual sources from Ancient Arabia, which also exposes its records in standard formats (oai_dc, EpiDoc, EDM) in an OAI-PMH repository. The initiatives of reuse of DASI open data in the frame of the recently ANR-funded project Maparabia (CNRS-CNR) are discussed in the paper, focusing on the exploitation of DASI's onomastic and geographic data in a new reference tool, the Gazetteer of Ancient Arabia. After introducing DASI and Maparabia projects and highlighting the objectives of the Gazetteer, the paper describes the conceptual model of its database and the module importing data from DASI. The population of the Gazetteer, implying also a data entry and manipulation phase, is exemplified by the case-study of the Ancient South Arabian place 'Barāqish/Yathill'. Based on the above experience, limitations and opportunities of data reuse and synchronisation issues between systems are discussed.

I corpora annotati, a condizione che adottino standard internazionali ed esponano i dati in formato aperto, hanno molte più possibilità, rispetto ai corpora tradizionali e analogici, di essere riutilizzati per obiettivi diversi da quelli per cui sono stati concepiti. Il presente articolo intende presentare i risultati di una precoce adesione alle buone pratiche e ai principi successivamente codificati come Open Science e FAIR nell'ambito di progetti di corpora testuali digitali,

specificamente in un campo di ricerca che possiamo definire di nicchia, ovvero l'epigrafia dell'Arabia preislamica. Il caso di studio analizzato in questo articolo è il Digital Archive for the Study of pre-Islamic Arabian inscriptions - DASI, un corpus online annotato delle fonti testuali dell'Arabia antica, che espone i suoi record anche in formati standard (oai_dc, EpiDoc, EDM) in un repository OAI-PMH. L'articolo presenta le iniziative di riuso dei dati onomastici e geografici di DASI in un nuovo strumento di reference, il Gazetteer of Ancient Arabia, sviluppato nel quadro del progetto Maparabia (CNRS-CNR) recentemente finanziato dall'ANR. Dopo un'introduzione ai progetti DASI e Maparabia, in cui sono esposti gli obiettivi del Gazetteer, l'articolo descrive il modello concettuale del suo database e il funzionamento del modulo di importazione dei dati da DASI. Il popolamento del Gazetteer, che implica anche una fase di inserimento e manipolazione dei dati, è esemplificato dal caso di studio del sito sudarabico di 'Barāqish/Yathill'. Tale esperienza offre un'occasione per discutere delle limitazioni e delle opportunità di riutilizzo di dati e metadati testuali, e delle questioni relative alla sincronizzazione fra sistemi.

Introduction

Building large, annotated textual corpora is hard-working, time consuming and expensive. Moreover, work and skills gained by researchers committed to this activity are still scarcely acknowledged by the academic community. However, annotated corpora, provided that they adopt international standards and best practice, and expose data in open format, have many more chances to be easily exploited and reused for different objectives than traditional, analogue corpora ([4]). This paper illustrates this thesis by focusing on the potential of the open-access archive of pre-Islamic Arabian inscriptions DASI, which is the main source of data for a different and further reference tool, the Maparabia digital Gazetteer of Ancient Arabia.

The first part of the article presents the project Maparabia and the DASI archive. The central paragraphs describe the Gazetteer system, i.e. its conceptual model and architecture, critically discussing opportunities and limitations in the reuse of DASI data within the Gazetteer. The case-study of the ancient South Arabian place of 'Barāqish/Yathill' exemplifies the functioning of the tool and the choices operated in its design. The paper ends with a focus on the challenges faced and the solutions devised in the synchronization of data between DASI and the Gazetteer.¹

DASI archive: A source of onomastic and geographic data

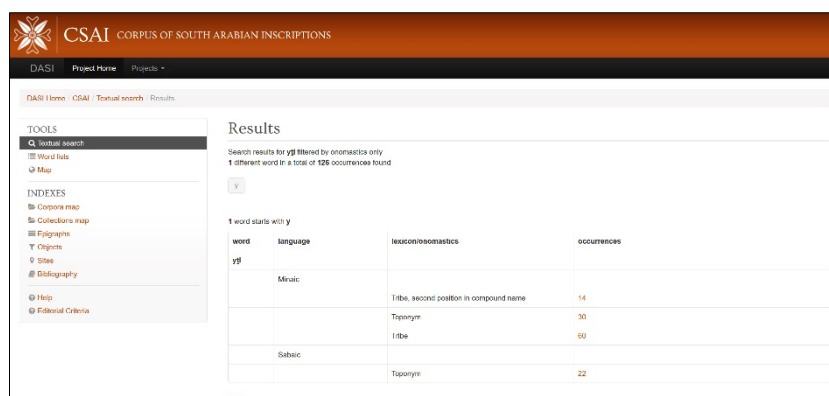
Over the past years, considerable advancements have been made in the research on Ancient Arabia, leading to the production of a mass of archaeological and textual data. The epigraphic databases have played an outstanding role in disseminating this knowledge through the web

1 This article is the result of a joint effort of all the authors. Only for academic purposes, part 4 (The Gazetteer of Ancient Arabia) is specifically attributed to A. De Santis, and part 5 (Case-study: Barāqish/Yathill) to I. Rossi.

publication of tens of thousands of annotated inscriptions and graffiti, spanning 1,500 years of history and covering a region that had long remained at the margins of the Near Eastern studies.

DASI - Digital Archive for the Study of pre-Islamic Arabian Inscriptions² is an online archive publishing at present the curated edition of nearly 8,500 inscriptions from Ancient Arabia ([1]). DASI corpus is for the most part made of the epigraphic heritage of the Ancient South Arabian civilization, which flourished since the early 1st millennium BCE until the advent of Islam, in the region corresponding to modern Yemen and neighbouring areas (the classical *Arabia Felix*) ([3]). The Ancient South Arabian textual heritage is composed by over 10,000 inscriptions and graffiti, and hundreds of texts on wooden sticks ([11]).

The information encoded in DASI primarily regards the epigraphic texts. Annotation of the textual phenomena is made according to the TEI-EpiDoc schema ([6]) in an XML encoding module embedded into the database. In DASI, the textual annotation concerns transcriptional, philological, and linguistic information – with a focus on onomastics. Names are essential components of an epigraphic source as they provide direct information on individuals and historical personalities, on the components of society, on divine figures, on places of the natural and anthropic space, etc. Names, as well as lexicon and portions of text, can be searched for in DASI website through the Textual Search tool. The Word List, a general index of the words, is also automatically generated in DASI, allowing to list all lexical and onomastic occurrences. Names can be classed by specific onomastic type (Figure 1). The Word List currently indexes more than 7,000 name-forms, with a total of over 45,000 occurrences in the Ancient South Arabian languages, comprising: personal names; divine names; names of months and decades; names of objects; names of buildings; names of social, political, and geographical entities. The latter two categories presently include about 3,300 names.



word	language	lexis/onomastics	occurrences
ytł	Minaic	Tilbe, second position in compound name	14
		Tporynre	30
Sabaic	Sabaic	Tilbe	60
		Tporynre	22

Figure 1: Summary of the occurrences of an onomastic item (ytł) in the Corpus of Ancient South Arabian Inscriptions within DASI.

- 2 DASI corpus is consultable through indexes and search tools at [http://dasi.cnr.it/]. DASI is the output of a five-year research project funded by the ERC from 2011 to 2016 (GA 269774; PI: Alessandra Avanzini, University of Pisa). It is currently maintained at the Consiglio Nazionale delle Ricerche.

DASI epigraphic records are also provided with a series of textual and contextual metadata, such as information on script and language, chronology, text genre, type of support and iconographic elements, archaeological and geographical context. The sites that are places of provenance (production or discovery) of the epigraphs are catalogued in specific records. More than 400 sites are indexed in DASI, and provide information about: ancient and modern toponymy; location (country, geographic area and present governorate, coordinates and related accuracy); types of the findings, architectural structures and monuments; history and chronology; history of research; kingdoms, languages, deities and tribes attested at that site; general description; bibliography. Each site record may be linked to the other ones, thus representing the spatial relations between them.

The Maparabia project: Analysing and synthesising information related to the territory of Ancient Arabia

The opportunity to exploit and enhance this wealth of geographic data and of onomastics having relation to a territory in Ancient Arabia, has occurred with the Maparabia project,³ which aims at developing tools for their analysis, producing syntheses, and making them both accessible to the greatest number.

Based on archaeological data ([8]) and large epigraphic corpora (DASI, OCIANA),⁴ the project has three main milestones or research instruments, freely accessible online, and adhering to Open Science and FAIR principles.

The first is a Digital Atlas of Ancient Arabia. This online platform is designed for the mapping of monuments, inscriptions, languages, scripts, cults and social groups. It exposes data of a geolocated database (PostgreSQL) of two types of entities: archaeological sites and inscriptions. The archaeological data is compiled from the existing bibliography; the epigraphic data is imported from the DASI corpus (about 7,200 South Arabian inscriptions at the moment) and OCIANA (about 3,500 North Arabian inscriptions). The platform includes full web-GIS functionality. Names of deities and tribes encoded in the DASI archive (see above) allow the mapping of the distribution of cults and social groups. Any specialist wishing to carry out advanced queries or spatial analyses may request access to the QGIS project and use the database on open-source GIS software for his/her own research.

3 Maparabia is a 5-year project funded by the French National Research Agency (ANR-18-CE27-0015), which encompasses a variety of fields of research: history, archaeology, epigraphy, linguistics, palaeography, geomatics, and geography. The fifteen members of the project come from four laboratories dedicated to pre-Islamic Arabia: CNRS-Orient & Méditerranée (Paris), CNRS-Archéorient (Lyon), Dipartimento di Civiltà e Forme del Sapere of the University of Pisa, and CNR-ISPC (Milan) [<http://www.orient-mediterranee.com/spip.php?article4002>].

4 OCIANA - Online Corpus of the Inscriptions of Ancient North Arabia [<http://krcfm.orient.ox.ac.uk/fmi/webd/ociana>].

The second is a *Thematic Dictionary of Ancient Arabia (TDAA)*. This is a comprehensive electronic dictionary covering several aspects of the history, society, religion, linguistics and topography of Arabia from the beginning of the first millennium BCE to the seventh century CE. The treatment of its entries is based primarily on the consideration of epigraphic and archaeological material, which in turn allows for an assessment of the reliability and historicity of other sources, often considered distorted by distance (classical sources) or remoteness (Arab-Islamic sources). In this respect, the *TDAA* is intended to meet the need for an up-to-date open-access synthesis. The *TDAA* will benefit both specialists of Ancient Arabia, and the wider academic community, by facilitating access to sources essential for the understanding of the ancient world and the emergence of Islam.

The third is a Gazetteer of Ancient Arabia. The Gazetteer consists of a list of places, providing their identification, description, and semantic relationships among them. The Gazetteer has been conceived to provide a complementary, semantic approach to the GIS, as it better points out the information about past geography provided by ancient texts, which is in the form of names, and allows to express cultural phenomena, such as political and administrative entities, which are not easily represented in their physical extension, and their numerous changes over time ([10]). Moreover, as gazetteers enhance the name-based search of spatial information and the spatially-oriented search of textual information on the web, which has a semantic organization, it is expected to support the description, discovery, understanding, and process of data about Ancient Arabia on the web ([9]).⁵

The Gazetteer of Ancient Arabia

The Gazetteer is designed to build upon several archaeological and geographical databases and textual archives, which have joined the project Maparabia or expose their data under open licenses. The aim is to organize and cross the information they include, and to stimulate study and reflection on fundamental research topics of Ancient Arabia that concern territorial dynamics, such as the settlement process and the man-environment relationship, and the socio-political organization.

Conceptual model and system architecture

The main entity of the Gazetteer of Ancient Arabia's conceptual model is the Place (Figure 2). The Maparabia Gazetteer adopts the definition of 'place' disseminated by the project Pleiades⁶ ([5]), therefore it takes into consideration elements of the natural and anthropic landscape, entire settlements and individual artifacts, political, social and cultural entities related to the territory,

5 The Gazetteer of Ancient Arabia is not yet openly available. It will be made public from 2022 onwards at [<https://ancientarabia.huma-num.fr/gazetteer/>]. Meanwhile, access to the database is allowed to authorized users at [<http://ancientarabia.cnr.it/gazetteer/>].

6 See the technical documentation available on the Pleiades website [<https://pleiades.stoa.org/help/concepts>].

‘whether or not exactly locatable, whether or not their actual relation with the real world can be ascertained’. Each Place record identifies univocally and persistently an ancient place and is related at least to one Location, that is its geographical expression, or one Name, that is its onomastic occurrence in an ancient written source. The relation with a Location or alternately to a Name is the condition of existence of a Place. A Place may have a physical or cultural-historical relation with another Place, that can be also chronologically qualified. Relations Location to Place, Place to Place and Place to Period are qualified by the degree of confidence of the associations.

Locations may be provided with Bibliography. Names have a link with at least one Source which witnesses its existence. This assumption, i.e. the acknowledgment of the source of a name, is central to the model of this Gazetteer and marks the difference from other similar projects.

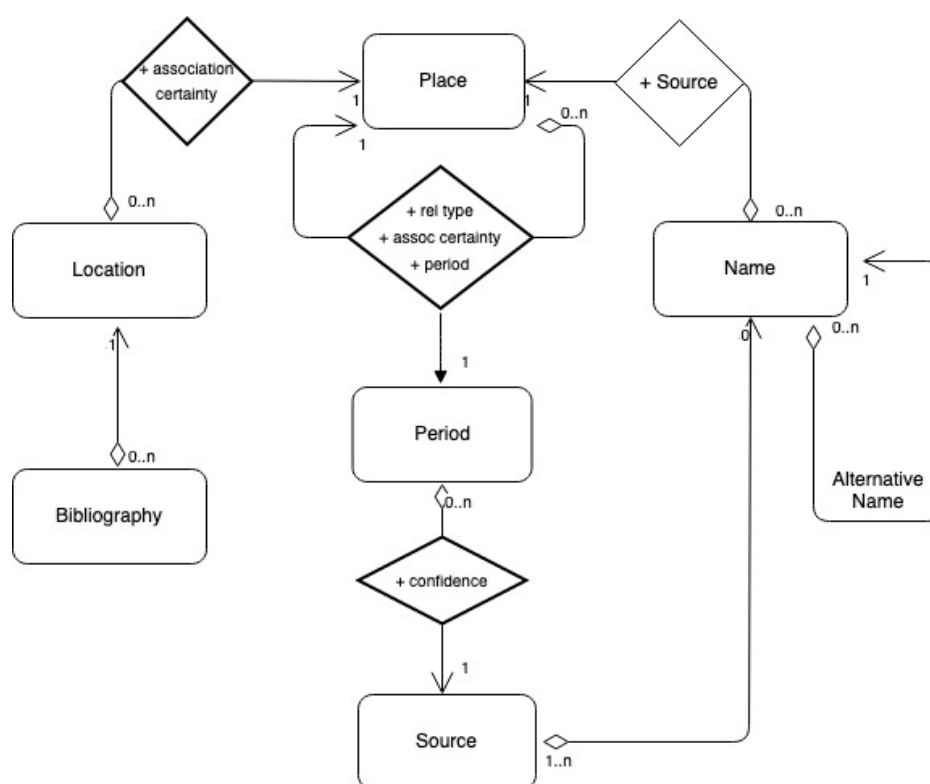


Figure 2: Conceptual model of the Gazetteer of Ancient Arabia.

The main source of place names of ancient Arabia, in particular of ancient South Arabia, is undoubtedly the huge, local epigraphic heritage. As mentioned above, inscriptions – mainly the ‘monumental’ inscriptions – put great emphasis on detailing the names of elements of the natural, anthropic, and social landscape, such as territories, mountains, rivers, settlements, cultivated fields, hydraulic facilities, public monuments and private buildings (even single parts or rooms of buildings), kingdoms, tribes and families.

Because of the availability of thousands of already annotated named entities and of data on archaeological sites, the bulk of data is to be imported from different systems, above all DASI. To this aim, the web application of the Gazetteer of Ancient Arabia includes a module importing data from the DASI web APIs. On the other hand, a data entry module has been developed in order to enrich with editorial content data already imported into the database and to include information that has not yet been encoded. For instance, Names and Sources, in particular those pertaining to the Arab-Islamic tradition, which are rarely available in digital format, can be added through the data entry module. A module exporting data completes the system architecture, allowing to expose Place records, being the other entities nested within, in JSON-LD format.

Importing data from DASI

Records and portions of texts are the main elements extracted from DASI system, which is a hybrid system (see the paragraph on the DASI archive, [2]). Prior to the import of data from DASI, an accurate mapping of fields has been performed. Importing criteria are adjusted to the type of Places occurring in DASI: archaeological site corresponding to an ancient settlement; modern site not identifiable with a place mentioned in ancient sources; monument related to an archaeological site; territorial entities mentioned in epigraphs.

As regards DASI records, the following ones are those imported into the Gazetteer:

- Sites, mapped into the Location records. One Site corresponds to one Location, whereas Monuments related to Sites are mapped in separate Locations. Location items cannot be modified; therefore, users must perform changes to content in DASI and then run a new import work. One Place record is automatically created for each Location record; whereas the place type is filled out according to the mapping criteria (Table 1), the remaining fields must be enriched, as the automatic creation is aimed at the first population only.
- Epigraphs, mapped into Source records. Each Name inherits the relation with the epigraphs which it occurs in. Sources are not editable, but new records can be added.
- Periods are imported into the vocabulary of the same name. Furthermore, relations between Sources and Periods are automatically created, according to the chronology of the Epigraph records of DASI. Periods can be changed and added, and relations with Sources and Places can be multiplied, as many are the (historical, linguistic, cultural, etc.) parameters that can be taken into consideration when dating inscriptions and places. However, users have the task to verify the coherence of chronologies, so that relations are consistent.

Place type	DASI Site / Name type
archaeological site	DASI Site > ancient name = unknown AND modern name ≠ unknown
building <ancient>	Name type = building / sacred place

environmental element	
geographic area <modern>	DASI Site > ancient name = unknown AND modern name = unknown AND region ≠ unknown
kingdom	
monument <modern>	DASI Site > monument
settlement <ancient>	DASI Site > ancient name ≠ unknown
social / administrative entity <ancient>	Name type = tribe Name type = nisbe

Table 1 Mapping of the Place type for automatic creation of Place records

As regards the textual elements, the Gazetteer retrieves items from the DASI Word List, limited to those enclosed by <placeName><placeName> and <orgName></orgName> tags, plus the <rs></rs> tag encoding the nisbe, i.e. the adjective describing a relation with territorial and/or social entities (e.g. Roman, Italian, etc.). Each pair, onomastic item plus EpiDoc compliant tag of the DASI Word List, is used to create one instance of the entity Name of the Gazetteer (Table 2). This inherits also the relation with the DASI epigraphic source it is attested in, just as Source inherits the Period it is dated to in DASI.

DASI EpiDoc compliant tag	Name type
<placeName></placeName>	toponym
<placeName type="sanctuary"></placeName>	sacred place
<placeName type="building"></placeName>	building
<rs type="nisbe"></rs>	nisbe
<orgName type="tribe"></orgName>	tribe

Table 2 Values of the field ‘name type’ populated from DASI

Case-study: Barāqish/Yathill

The example of the place ‘Barāqish’, ancient *ytʿl* (conventionally vocalised Yathill), illustrates the conceptual model and the functioning of the Gazetteer’s data entry interface, as regards both the manipulation of DASI-imported data and the addition of new information.

Location

Barāqish, as an ancient South Arabian archaeological site located in north-western Yemen, is described by a Site record in DASI [http://dasi.cnr.it/sit-243]. Data imported from this record (country, coordinates, coordinates accuracy, type of site, structures, location and toponymy, history of research, general description and chronology) and from the linked Bibliography records of DASI, populate the Location record of the Gazetteer, that cannot be manipulated.

Name

The ancient city of Yathill is mentioned as *ytł* <placeName> in 25 Minaic inscriptions encoded in DASI for a total of 30 occurrences, and in 11 Sabaic inscriptions for a total of 22 occurrences (Figure 1, Figure 3).

language	epigraph	context
Central Minaic	as-Sawdā' 13	» Hmryj) w-bn s'hm Ytł gn' b'k'rbj(b'k'rbj) w- «
Central Minaic	Gr 326	» S' [...] Ytł [...] b'yn (g'ly s'hm Ytł w-s'hm M n qdm «
Central Minaic	M 172	» d- hm h(ł) s' hgm Ytł b'q(m) [...] Ytł [...] s'rbt «
Central Minaic	M 176	» [...] Ytł m rb Ytł [...] M'jm b- «
Central Minaic	M 177	» [...] Ytł [...] b-hgm Ytł n'mn w-hf w- «
Central Minaic	M 185	» s'hd b-gn' hgm Ytł kl'zwt m'k'n d- «

Figure 3: List (partim) of the occurrences of the toponym *ytł*, annotated as a <placeName>, in the Minaic inscriptions.

Based on this toponym, the Gazetteer is automatically provided with a corresponding Name record, which can be enriched in the Gazetteer with comments and information regarding the language(s) of its occurrences, the accuracy and completeness of transcription, and links with alternative appellations, such as different spellings of the name (Figure 4).

Name-to-Name relation

This is not the case of *ytł*, but heterography frequently occurs in the Semitic inscriptions, for instance in words with a long vowel (*mater lectionis*) which can be spelled in both their *scriptio plena* and *defectiva* (e.g. *Mrb* vs *Mryb* for the ancient name of the Sabaeen capital). Alternative appellations are managed at the Name-to-Name relation level, in that the two Name records are associated to each other as alternative names.

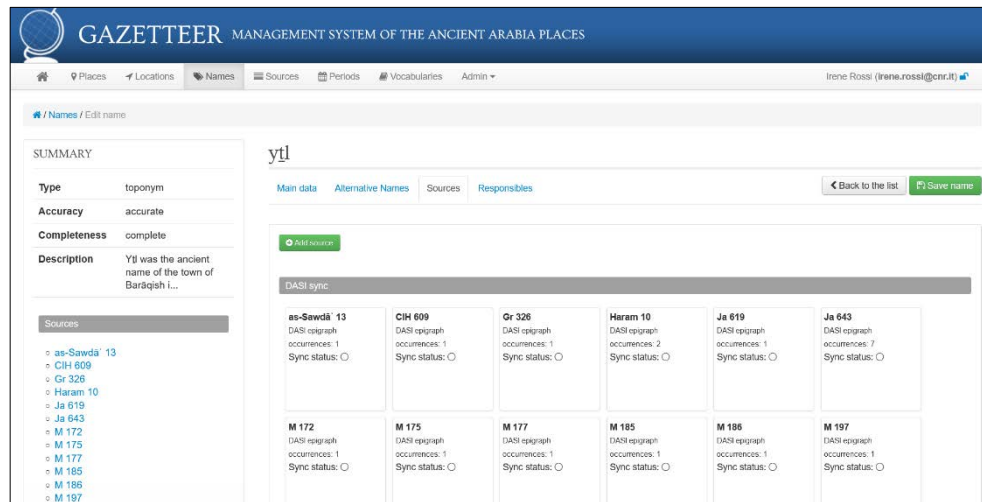


Figure 4: Name record 'ytl' in the Gazetteer, showing the relations with the Sources attesting the name.

Source

For each inscription featuring at least one occurrence of the toponym *ytl*, the Gazetteer is automatically enriched with a Source record (e.g. Gr 326, Figure 5), deriving from DASI the relations with the Names it attests, besides information such as the epigraphic siglum, the chronology (Period) and the URI of the corresponding DASI record [http://dasi.cnr.it/csai-epi-7547]. As one inscription may feature more than one onomastic item concerned with the Gazetteer, each Source record in the Gazetteer is linked to the Name records of all the relevant named entities its text contains. For instance, Gr 326 also mentions the toponym *m'n* (modern Ma'in, the ancient capital of the Minaean kingdom also governing Yathill; see 'Linked names' in Figure 5).

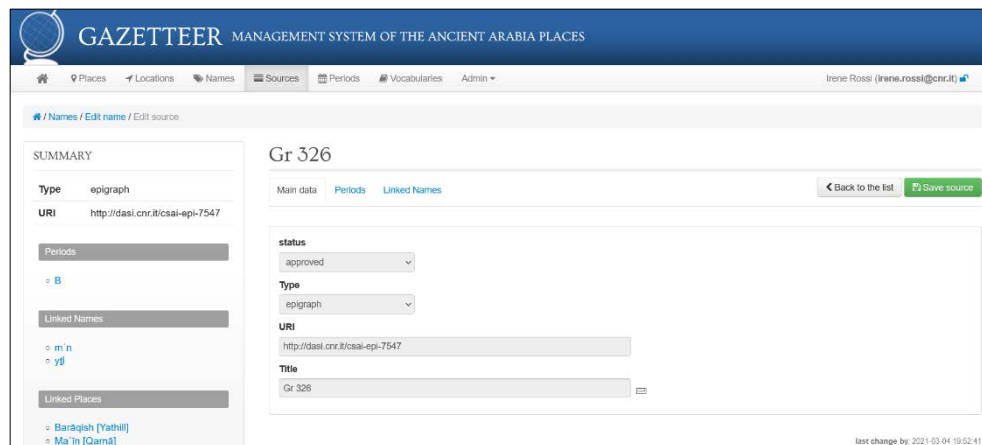


Figure 5: Main tab of the Source record 'Gr 326' in the Gazetteer, linked to the Name record 'ytl'.

Period

Chronological information on each inscription is described by means of the Source's relation with one or more Period records (see Figure 5: 'Periods'). These relations are imported from DASI, where a chronological grid for the Ancient South Arabian corpora, organised in broad periods, is applied to the inscriptions (Period 'B' roughly spans the second half of the first millennium BCE). However, the Gazetteer also allows to manually enter new Period records and to relate more than one Period to one Source. To the aims of the Gazetteer, this is useful for instance when a source is datable to a narrower chronological range with respect to DASI grid's periods, or to a different periodization, especially if the Source record itself is not automatically imported from DASI but created from scratch in the Gazetteer as in the case of Classical sources.

New records

In fact, one essential feature of the Gazetteer is the possibility of entering new Name, Source and Period records, besides those imported from DASI. This makes the Gazetteer an open system: on the one hand it draws from an annotated corpus such as DASI all the already available information, avoiding manual re-entering of data, but on the other hand it remains expandable beyond the limits of its main source corpus. This allows to record the names of the geographical entities comprised within the scope of the Maparabia Project, be they attested in further linguistic corpora of inscriptions (from Arabia and beyond, e.g. Greek, Latin, Aramaic, North Arabian etc.), or in other kinds of sources, above all Classical and Arab-Islamic literature. For instance, a new Name record was created in the Gazetteer, dealing with Yathill's Greek toponym Ἰαθουλα, derived from Strabo's *Geography* (16, 4, 24). The latter constitutes in its turn a newly entered Source, assigned with a proper Period (1st century BCE).

Place

As the archaeological site of Barāqish corresponds to the ancient city of Yathill, both the Location and the Name records are connected to the Place Barāqish [Yathill]. Place is fully editable in the Gazetteer, precisely because it is the main container of new information within the system. It can be assigned one or more types (e.g. Barāqish is both an archaeological site and an ancient settlement), it can be provided with external matches in the case it is recorded in other gazetteers (e.g. Pleiades for the Ancient World: <https://pleiades.stoa.org/places/39300>), and it can be textually described in order to summarise its main features (Figure 6).

The screenshot displays the 'Gazetteer Management System of the Ancient Arabia Places' interface. The main content area is titled 'Barāqish [Yathill]'. On the left, a 'SUMMARY' sidebar lists various attributes: Type (archaeological site settlement <ancient>), Description (Barāqish, ancient Yathill (yī), is an archaeologi...), Locations (Barāqish (certain)), Related Names (yī (toponym), yīy (nisba), ʾAḥḥouka (Aḥrouka) (toponym)), Related Places (includes Barāqish - Temple of ʾAḥtar dhi-Qaḥd (certain) - Period: Pre-Islamic period, east of Barāqish - Neropolis (certain) - Period: Pre-Islamic period, includes Barāqish - Temple Barān of Nakrah (certain) - Period: Pre-Islamic period, includes Barāqish - City Walls (certain) - Period: Pre-Islamic period, has attestation of yī / d yī (certain) - Period: B), Responsibilities (Irene Rossi, Alessandra Lombardi), and Rights (CNR - CNRS). The main content area includes a 'status' dropdown (approved), a 'Type' dropdown (archaeological site), a 'Title' field (Barāqish [Yathill]), a 'Description' text area, a 'Period' dropdown (Pre-Islamic period), a 'Provider' field (Gazetteer of Ancient Arabia - CNR/CNRS), a 'Publication date' field, and an 'External match' field (https://pleiades.stoa.org/places/39300). A 'last change by: Irene Rossi / 2021-10-16 16:26:29' timestamp is visible at the bottom right.

Figure 6: Main tab of the Place record 'Barāqish [Yathill]' in the Gazetteer.

Place-to-Name relation

By means of relations, the Place aggregates the other relevant kinds of data registered in the Gazetteer, which complete its description. The link with one or more Location and Name records can be enriched with information on the association certainty. Each Name carries the assumed relation with the Sources it is mentioned in and, consequently, the Period of its attestation, thus providing the complete spectrum of the variations in the onomastics of a given Place, depending on the language and time of attestation.

One central challenge in the design phase of the system was the management of homograph names in what concerns their relation to Places. Homograph names of sites and tribes are not extremely frequent, but there is a considerable number of them even in the Ancient South Arabian corpus alone. Homograph names of elements of the natural environment or of buildings or sacred spaces are even more common. For instance, there is proof that there were at least two sanctuaries of the god Almaqah named Awwam (*'um*): one was the main Ancient South Arabian confederal sanctuary in the surroundings of the capital Ma'rib, the other one was located in a peripheral area of the Sabaeen kingdom. Another example regards the several tribes named Bakilum: in order to distinguish among themselves, already in ancient times a specification indicating the city of origin was sometimes added to the name *bklm* (e.g. *bklm d-mrymtm*, the 'Bakilum of (the city of) Maryamatum'), but this is not always the case. Each pair 'name-form +

onomastic tag’ in DASI Word List is imported in the Gazetteer as a distinct Name; this means that the same name-form which is marked-up by different tags (e.g. the toponym *yṯl* <placeName> and the name of tribe *yṯl* <orgName type="tribe">, Figure 1) is treated as two separate Name records. On the other hand, DASI mark-up does not disambiguate between homograph names having the same tag; the identification of named entities is precisely one of the aims of the Gazetteer. In order to do so, a functionality of the interface was implemented, which allows to dissociate specific Sources of one Name in the latter’s relation with a Place, by de-selecting the irrelevant Source entries (e.g. the inscription CIH 609 in Figure 7).



Figure 7: Management of Place-to-Name relations in the Gazetteer, with (de)selection of non-pertinent Sources.

Place-to-Place relation

The relation Place-to-Place allows to draw links between ancient settlements having, for instance, a relationship of geographic proximity; moreover, it records the ‘is part of/includes’ relationships between a Place of type ‘archaeological site’ and a Place of type ‘monument’. This kind of relation is especially relevant as most of these monuments are mentioned with their names in the epigraphic sources, such as the temple of the god Nakraḥ in Barāqish, named Barān (*brn*) (Figure 8).

Another kind of Place-to-Place relation is the one between the Place of type ‘settlement’ and the Place of type ‘social / administrative entities’. For instance, *yṯl* ‘Yathill’ (with its alternative name *d yṯl* meaning ‘the one of Yathill’) is also the appellation of the tribe (<orgName type="tribe"></orgName>, Figure 1) based in the city of Yathill; therefore, a Name and a Place records are created for this tribe. The Place record of the tribe *yṯl* / *d yṯl* (type: social / administrative entity) can be linked to the Place Barāqish [Yathill] (type: settlement) (Figure 8), and to any further ancient settlements it might have been connected with.

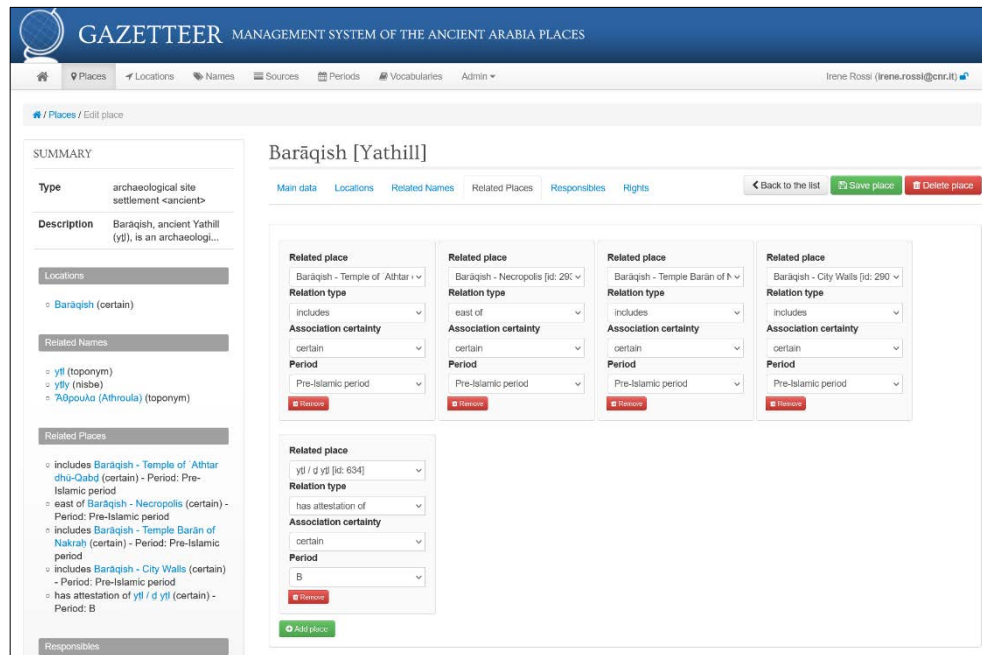


Figure 8: Management of Place-to-Place relations in the Gazetteer, including relations with Place-types 'monument' and 'social / administrative entity'.

Issues in data reuse

The dependence from other repositories clearly underlines that the Gazetteer is conceived as a tool of reuse, exploitation and enhancement of sources, in particular textual sources, already digitized.

Limitations and opportunities

Precisely the selection of the marked elements has brought out issues regarding the reuse of encoded texts. According to the encoding criteria established for DASi, individual onomastic elements can occur within onomastic groups: eponyms (<rs type="eponym"></rs>), compound names (<rs type="complex"></rs>), nominal groups (<rs type="nominalGroup"></rs>), and signatures (<seg type="signature"></seg>). DASi Word List indexes single onomastic elements regardless of the onomastic groups they are embedded in. This poses problems, in particular, in compound names whose aim is to define unique onomastic entities without distinguishing single components.

There is a scientific issue, underlying this problem, that has no solution in itself. The objective of DASi project, indeed, was the achievement of the digital edition of the pre-Islamic Arabian epigraphic heritage. Moreover, the linguistic study of the inscriptions has been one main aim of

the encoding process since the first phases of the project, going back to the end of the 1990s.⁷ The Gazetteer, instead, intends to exploit those digital editions for specific research aims, one of them being the study of ancient toponyms. Since no encoding work can envisage an unlimited set of phenomena to be codified for any possible future use, technical solutions must be devised.

For instance, the site in Saudi Arabia now called Qaryat al-Fāw received several appellations in the Ancient South Arabian inscriptions, all to be referred to the same place:

- *qryt*: <placeName>qryt</placeName>
- *qrytm*: <placeName>qrytm</placeName>
- *qryt ṭlw*, probably to be interpreted as ‘Qaryat the Red’: <rs type="complex"><placeName n="1">qryt</placeName> <placeName n="2">ṭlw</placeName></rs>
- *qrytm ḏt khlm*, lit. ‘Qaryat that of Kahl’ (Kahl being the local god): <rs type="complex"><placeName n="1">qrytm</placeName> <placeName n="2">ḏt</placeName> <placeName n="3">khlm</placeName></rs>.

This results in *qryt*, *qrytm*, *ḏt*, *khlm* and *ṭlw* being indexed as separate onomastic items corresponding to each of the <placeName> elements in DASI Word List,⁸ and would therefore generate separate Name records in the Gazetteer. While this processing of words was based on DASI’s interest towards linguistic annotation, the Gazetteer has the different objective of identifying and disambiguating the geographic/territorial entities designated by specific names, thus associating the names *qryt*, *qryt ṭlw*, *qrytm* and *qrytm ḏt khlm* with one single Place entity.

To overcome this idiosyncrasy between the two systems, the Gazetteer import module has been designed so that compound names including <placeName></placeName> and <orgName></orgName> tags are reconstructed and mapped into Name records. After a survey of the compound names encoded, a set of conditions the import module must check have been established. Since each text is considered a sort of matrix and each word has a position defined by the intersection between a line and a column, only onomastic elements that comply with all of them are aggregated in a compound name:

- they are attested in the same inscription,
- at the same line

7 The focus on the linguistic description of the texts and the pioneering venture of the project in the realm of the digitization of Ancient Arabian epigraphy informed the choices made in the original encoding of the onomastic phenomena. This has sometimes brought about legacy issues with which it was not always possible to cope within the scope of DASI project. For a detailed description of DASI encoding and the discussion of such limitations, see [2].

8 Going into deeper detail, the occurrences of the entry (word-form) *qryt* in the Word List are grouped into sub-entries: one described as a ‘toponym’ (for the occurrences of *qryt* alone) and one as a ‘toponym, first position in compound name’ (for *qryt ṭlw*, and potentially for any other compound toponym having *qryt* as its first element). The case of *qrytm ḏt khlm* is even more complex, as one of its <placeName> elements, namely *ḏt*, also occurs in other toponyms, such as *ḏt ḡylm* (modern Hajar Ibn Ḥumayd in Yemen).

- in linear sequence (order of the occurrence)
- they belong to the same name type
- they are placed on consecutive columns (except for clitics that have the same position of the word they are connected to).

Synchronization issues

The Gazetteer architecture, in that it depends from a different system, has raised also synchronization issues. More generally, synchronization and versioning are issues to be faced when reusing data.

The Maparabia gazetteer has a one-direction synchronization: changes to contents in the Gazetteer do not turn back to DASI, whereas any change to DASI may cause updates in the Gazetteer, by addition or rewriting, in particular to editable fields. The choice of not simply exporting data from DASI into the Gazetteer *una tantum*, but of synchronizing the two systems is due to the consideration that DASI, which is the base source of the Gazetteer's data, is not a closed project: not only are corpora constantly updated with new epigraphic material, but editions are also corrected on the basis of improved readings and new linguistic and cultural interpretations. We deemed it essential that the evolving life of DASI data could implement the Gazetteer, in order to make of it an always up-to-date tool.

In order to achieve this goal, the Gazetteer keeps tracks of the updates to the contents and allows to control the results of the import works, by excluding records from updates. There is no proper control panel, but this is distributed among the sections corresponding to the Gazetteer entities. Indeed, each section is provided with a list of previews that users can filter and browse by sync status. The list is updated when an import work is accomplished and shows changes that were performed on DASI and are likely to affect contents of the Gazetteer: creation and deletion of onomastic items of the Word List and Site records, but also modifications that cause changes in Names, Locations and their relations; creation and deletion of Epigraph records that affect Sources; amendments to the encoded texts that modify occurrences of onomastics; creation, modification and deletion of Periods.

This display method replaces notifications. Moreover, changes are finally transferred into the Gazetteer, only after a manual validation is carried out. Each record preview is provided with a circle, the icon that represents the sync status, and a button to be clicked for validation. The colors of the circle are codified according to the sync status:

- green: a new record was created in DASI and has just been imported into the Gazetteer. Users, once the record itself is checked and possible relations are evaluated, are invited to validate import. Thus, the circle becomes empty
- red: the record has been deleted in DASI, but is still present in the Gazetteer. Once users have checked its relations, it can be definitively deleted by the basket icon
- blue: the corresponding record in DASI has been recently updated. Users are required to take note and validate the update; afterward the blue circle becomes empty
- empty circle: no action is required.

Users have therefore active role: since no change becomes permanent without their intervention, they are stimulated to enrich contents by integrating Name files, creating new Places and relations among entities.

Not only records are provided with sync status. The relation Name to Source, namely the occurrences of names in a particular inscription, are subject to synchronization: if even one of the occurrences is modified, the overall status changes according to the color codification above. Validation, however, is not required. It is aimed only at drawing the attention to changes and their potential consequences.

For instance, if a new inscription attesting ns^2n <toponym> (the ancient name of the site of as-Sawdā' in northern Yemen, conventionally vocalized Nashshān) is added in DASi, an import work will create a new Source record (green) in the Gazetteer for that inscription. As the Name record ns^2n <toponym> was already present in the Gazetteer because it is already attested from other sources, it will appear in the list of Names provided with a blue circle (Figure 9): this indicates that a synchronization change has happened to the Name record.

Title	Type	Languages	Accuracy	Completeness	Status	Last change by	Id	DASI sync
1. na'n	toponym	Ancient South Arabian	accurate	complete	approved	Irene Rossi / 2021-02-05 09:22:20	101	●
2. ns'n	tribe	Ancient South Arabian	accurate	complete	approved	Irene Rossi / 2021-02-08 11:55:38	219	○
3. ns'nytr	nisbe	Ancient South Arabian	accurate	complete	approved	Irene Rossi / 2020-10-01 16:42:03	1628	○

Figure 9: Example of change in the circular icons' colours, reflecting synchronization updates.

This change precisely regards the addition of a new Name-to-Source relation, which will be indicated by a green circle in the 'Sources' tab of the ns^2n <toponym> Name record; the 'old' (i.e. already approved) Name to Source relations feature an empty circle instead (cf. e.g. Figure 4). The user is called to validate the changes by means of the dedicated button. The addition of Sources has also consequences with respect to the Place-to-Name relation, as the user has to confirm that the name-form in that specific Source is correctly attributed to that Place (the automatic selection being disabled because of potential homograph Names; cf. Figure 7).

If, instead, one of the 'old' Name-to-Source relations is broken because, on the basis of a new reading or interpretation, the spelling of the only one occurrence of the toponym in a specific inscription is changed (e.g. from the toponym [... n] s^2n to the nisbe [... 's²] s^2n 'the Nashshānites'), the Name-to-Source relation will feature a red circle. If that Source features instead further occurrences of ns^2n <toponym>, the circle will be blue.

Finally, if a name is only known by one occurrence in one DASI inscription (i.e. the Name record is linked to only one Source) and the name-form is changed in (or the name is completely deleted from) that inscription, not only the Name-to-Source relation will be displayed in red, but also the Name record itself. If the Gazetteer user agrees with such editorial changes in DASI, after having deleted the relevant relations between records, he/she will be also able to delete the Name record in the Gazetteer.

Conclusions

In conclusion, if the historical and cultural domain of the Maparabia Gazetteer – i.e. Ancient Arabia – is its main peculiarity, its added value, compared to the other gazetteers of the ancient world, is the direct and ‘living’ bond with the annotated epigraphic corpus providing it with the direct sources of ancient toponymy. The core data of the Gazetteer results automatically from the mass digitization of the direct written heritage of pre-Islamic Arabia conducted during the DASI project, according to guidelines that have established themselves as proper ‘standards’ in the digital epigraphy field, and applying the best practices that were subsequently formalized under the label of FAIR principles ([13]). DASI records, provided with URIs, are exposed in standard formats (oai_dc, EpiDoc, EDM) in an OAI-PMH repository, thus allowing different projects to access and use its data.

Standards and best practices that were observed have ensured the effective and easy reuse of texts, even though the scientific objectives of the Gazetteer are significantly different from the original purpose of the digitization and scope of use of the digital edition. Imported data can be enriched in the Gazetteer with additional details arising from the thorough study of the sources and the effort to systematize identification and relations. The instances of the main entity Place, in fact, are the only ones that must be created from scratch, requiring the editorial intervention to disambiguate, identify and circumscribe an ancient ‘place’. As well as the creation of the relation Place to Place, this is the step of the workflow the scientific reflection focuses on, and the editorial responsibility is more significant.

The export module of the Gazetteer allows to expose Place records, being the other entities nested within, in JSON-LD format. Each Place item is identified with a URI and is released under open license. This is consistent with the philosophy that has allowed the Gazetteer itself to be created and is expected to increase dissemination through aggregation and linking with further gazetteers, and therefore with archaeological, textual and geographic data, pertaining to different chronological and cultural contexts ([7];[12]).

Acknowledgements

The research leading to this paper received funding by the French National Research Agency in the frame of the project Maparabia (ANR-18-CE27-0015, PI J. Schiettecatte, 2019-23). We wish to acknowledge the contribution of Dr. Alessandra Lombardi to the records of the Gazetteer of Ancient Arabia pertaining to the case-study described in this article.

References

- [1] Avanzini, Alessandra, Annamaria De Santis, and Irene Rossi. 2018. 'Encoding, Interoperability, Lexicography: Digital Epigraphy through the Lens of DASI Experience'. In *Crossing Experiences in Digital Epigraphy: From Practice to Discipline*, by Annamaria De Santis and Irene Rossi, 1–18. Warsaw: De Gruyter Open Poland. <https://doi.org/10.1515/9783110607208-002>.
- [2] Avanzini, Alessandra, Annamaria De Santis, Daniele Marotta, and Irene Rossi. 2014. 'Between Harmonization and Peculiarities of Scientific Domains. Digitizing the Epigraphic Heritage of Pre-Islamic Arabia in the Project DASI'. In *Information Technologies for Epigraphy and Cultural Heritage. Proceedings of the First EAGLE International Conference*, edited by Silvia Orlandi, Raffaella Santucci, Vittore Casarosa, and Pietro Maria Liuzzo, 69–93. <https://doi.org/10.13133/978-88-98533-42-8>.
- [3] Avanzini, Alessandra. 2016. *By Land and by Sea: A History of South Arabia before Islam Recounted from Inscriptions*. Arabia Antica 10. Roma: «L'Erma» di Bretschneider.
- [4] De Santis, Annamaria, and Irene Rossi, eds. 2018. *Crossing Experiences in Digital Epigraphy: From Practice to Discipline*. Warsaw: De Gruyter Open Poland. <https://doi.org/10.1515/9783110607208>.
- [5] Elliott, Tom, and Sean Gillies. 2010. 'Digital Geography and Classics'. In *Changing the Center of Gravity: Transforming Classical Studies Through Cyberinfrastructure*, edited by Melissa Terras and Gregory Crane, 223–62. Piscataway: Gorgias Press. <https://doi.org/10.31826/9781463219222-013>.
- [6] Elliott, Tom, Gabriel Bodard, Elli Mylonas, Simona Stoyanova, Charlotte Tupman, Scott Vanderbilt, and et al. 2007-2017. 'EpiDoc Guidelines: Epigraphic Documents in TEI XML (Version 8)'. <http://www.stoa.org/epidoc/gl/latest/>.
- [7] Isaksen, Leif, Rainer Simon, Elton T.E. Barker, and Pau de Soto Cañamares. 2014. 'Pelagios and the Emerging Graph of Ancient World Data'. In *Proceedings of the 2014 ACM Conference on Web Science (WebSci '14)*, 197–201. New York, NY, USA: Association for Computing Machinery. <https://doi.org/10.1145/2615569.2615693>.
- [8] Schiettecatte, Jérémie. 2011. *D'Aden à Zafar. Villes d'Arabie Du Sud Préislamique*. Orient & Méditerranée. Archéologie 6. Paris: De Boccard.
- [9] Shaw, Ryan. 2016. 'Gazetteers Enriched: A Conceptual Basis for Linking Gazetteers with Other Kinds of Information'. In *Placing Names: Enriching and Integrating Gazetteers*, edited by Merrick Lex Berman, Ruth Mostern, and Humphrey Southall, 51–63. Bloomington: Indiana University Press. <https://doi.org/10.2307/j.ctt2005zq7>.
- [10] Southall, Humphrey, Ruth Mostern, and Merrick Lex Berman. 2011. 'On Historical Gazetteers'. *International Journal of Humanities and Arts Computing* 5 (2): 127–45. <https://doi.org/10.3366/ijhac.2011.0028>.
- [11] Stein, Peter. 2020. 'Ancient South Arabian'. In *A Companion to Ancient Near Eastern Languages*, edited by Rebecca Hasselbach-Andee, 337–53. John Wiley & Sons, Ltd. <https://doi.org/10.1002/9781119193814.ch18>.
- [12] Tupman, Charlotte. 2021. 'Where Can Our Inscriptions Take Us? Harnessing the Potential of Linked Open Data for Epigraphy'. In *Epigraphy in the Digital Age. Opportunities and Challenges in the Recording, Analysis and Dissemination of Inscriptions*, edited by Isabel Velázquez Soriano and David Espinosa Espinosa, 115–28. Oxford: Archaeopress.
- [13] Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. 'The FAIR Guiding Principles for Scientific Data Management and Stewardship'. *Scientific Data* 3 (1): 160018. <https://doi.org/10.1038/sdata.2016.18>.