Versione PROVVISORIA del contributo presentato al Convegno Annuale

The digital Gazetteer of Ancient Arabia: an example of reuse and exploitation of annotated textual corpora

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

[1] Independent researcher, Italy – annamaria.desantis(«»)unipi.it
[2] Senior developer, Italy – matteogal(«»)gmail.com
[3] CNR – Istituto di Scienze del Patrimonio Culturale, Italy – irene.rossi(«»)cnr.it
[4] CNRS – UMR8167 Orient & Méditerranée, France – jeremie.schiettecatte(«»)cnrs.fr

ABSTRACT

This paper aims at presenting the results of the early adhesion to the principles afterward codified as Open Science and FAIR principles in the frame of digital epigraphic projects in a niche area of research such as the pre-Islamic Arabian studies, to show how 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. 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.

KEYWORDS

Ancient Arabia; annotated textual corpora; digital epigraphy; digital gazetteers; digital reference tools; open data; reuse and exploitation of textual content; semantic web.

1. 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. 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.

2. DASI ARCHIVE: THE SOURCE OF ONOMASTICS 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 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 Inscriptions[1] is an online archive publishing at present the curated edition of nearly 8,500 inscriptions from Ancient Arabia[2]. The information encoded in DASI primarily regards the epigraphic texts. Annotation of the textual phenomena is made according to the TEI-EpiDoc schema [4] in an XML encoding module embedded into the database. The annotation concerns transcriptional, philological, and linguistic information – with a focus on onomastics. A general index of the names by specific onomastic type is automatically generated in DASI, allowing also to retrieve all their occurrences. It currently indexes more than 7,000 name-forms,

[1] DASI corpus is consultable through indexes and search tools at [http://dasi.cnr.it/] [1]. 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.

[2] 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*). The Ancient South Arabian textual heritage is composed by over 10,000 inscriptions and graffiti, and hundreds of textual sources on perishable material. See [2] and [3].

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.
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.

3. THE MAPARABIA PROJECT AND THE GAZETTEER OF ANCIENT ARABIA

The opportunity to exploit and enhance this wealth of geographic data and of onomastics having relation to a territory, has occurred with the MAPARABIA project[3]. The objective of MAPARABIA is creating several tools to integrate and perform analysis, with the aim of finally producing synthesis, on the considerable amount of datasets collected in the last 50 years of intensive research on the pre-Islamic history of the Arabian Peninsula. Besides producing a web-GIS, a historical atlas and an online thematic dictionary of Ancient Arabia, MAPARABIA envisages the creation of a further reference tool, the Gazetteer of Ancient Arabia[4].
This consists of a list of places, providing their identification, description, and semantic relationships among them. The MAPARABIA Gazetteer adopts the definition of "place" disseminated by the project Pleiades[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, "whether or not exactly locatable, whether or not their actual relation with the real world can be ascertained".
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 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, the socio-political organization, the pre-Islamic linguistic and writing landscape, the evolution of pre-Islamic religion and the origins of Islam.
Indeed, 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 [5]. 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 [6].

4. CONCEPTUAL MODEL AND SYSTEM ARCHITECTURE OF THE GAZETTEER

The main entity of the Gazetteer of Ancient Arabia's conceptual model is the Place. 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. Location may be provided with Bibliography; similarly, Name inherits a link with at least a Source which witnesses its existence (Fig. 1).

---

[3] MAPARABIA is a four-year research project funded by the French National Research Agency (project ANR-18-CE27-0015). It is coordinated by Jérémie Schiettecatte (CNRS-UMR8167 - Orient & Méditerranée), with the partnership of Mounir Arbach (CNRS-UMR5133 - Archéorient), Irene Rossi (CNR-ISPC) and formerly Alessandra Avanzini (University of Pisa). It involves a large number of researchers belonging to both French and Italian institutions [http://www.orient-mediterranee.com/spip.php?article4002].

[4] 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].

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

In this first phase of the project MAPARABIA, the Gazetteer is synchronized only with the system DASI. The web application consists of: a module importing data from the DASI web APIs, a database, a data entry and a module exporting data.

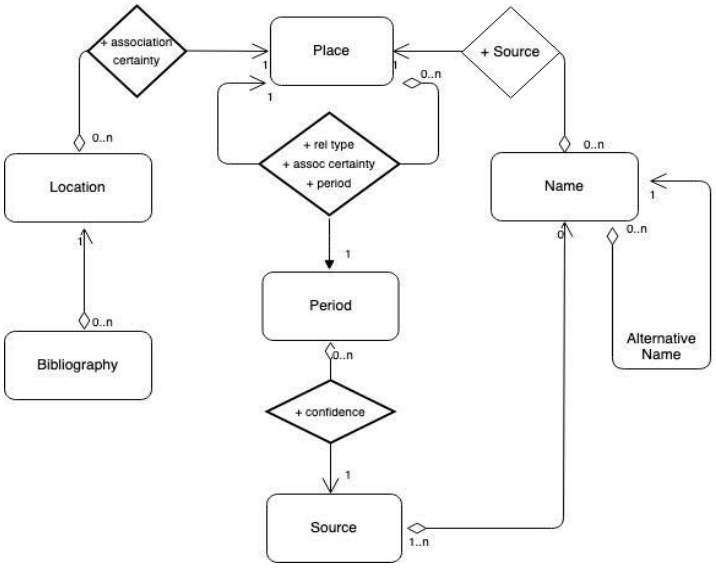

Figure 1. Conceptual model of the MAPARABIA Gazetteer

The population of the Gazetteer, indeed, is automatically performed by importing on demand data from DASI, in particular from:
- Site records (archaeological sites, which are provenance or place of production of inscriptions, and related monuments), implementing the Location records of the Gazetteer,
- Word List, limited to onomastics (toponyms, nisbas, tribe names, and names of (sacred) buildings attested by the inscriptions encoded), implementing Name records,
- Epigraph records, implementing Source records,
- and vocabularies, above all the Periods.

Each couple onomastic item + EpiDoc compliant tag of the DASI Word List is used to create one instance of the entity Name of the Gazetteer. It 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. Whereas Location records cannot be created or modified, Place, Name, Source and Period can be also entered manually in the Gazetteer and, when already existing, implemented with new metadata.

DASI EpiDoc compliant tag	**name type**
<placeName></placeName>	toponym
<placeName type="sanctuary"></placeName>	religious building
<placeName type="building"></placeName>	other building
<rs type="nisbe"></rs>	nisba
<orgName type="tribe"></orgName>	tribe

Table 1. Values of the field "name type" populated from DASI

The example of as-Sawdāʾ / Nashshān illustrates the conceptual model and the functioning of the Gazetteer. as-Sawdāʾ, as an archaeological site, is described by a Site record in DASI [http://dasi.cnr.it/sit-39]. 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.
This archaeological site corresponds to the ancient city of Nashshān, mentioned as *nsʾn* <toponym> in 2 Minaic inscriptions and 27 Sabaic inscriptions encoded in DASI (Fig. 2). The Gazetteer is then provided with a Name record, related to the Source records that correspond to the Minaic and Sabaic inscriptions of DASI, which can be enriched with comments and information regarding language, accuracy and completeness of transcription, and links with alternative

appellations, such as in different languages, different transliterations and so on. For instance, a new Name record can be created, dealing with the Latin toponym Nestum, which will be linked to *ns²n*, as Alternative Name, and to a new Source record dealing with Pliny the Elder's *Naturalis Historia* (VI, 32, 160) (Fig. 3).

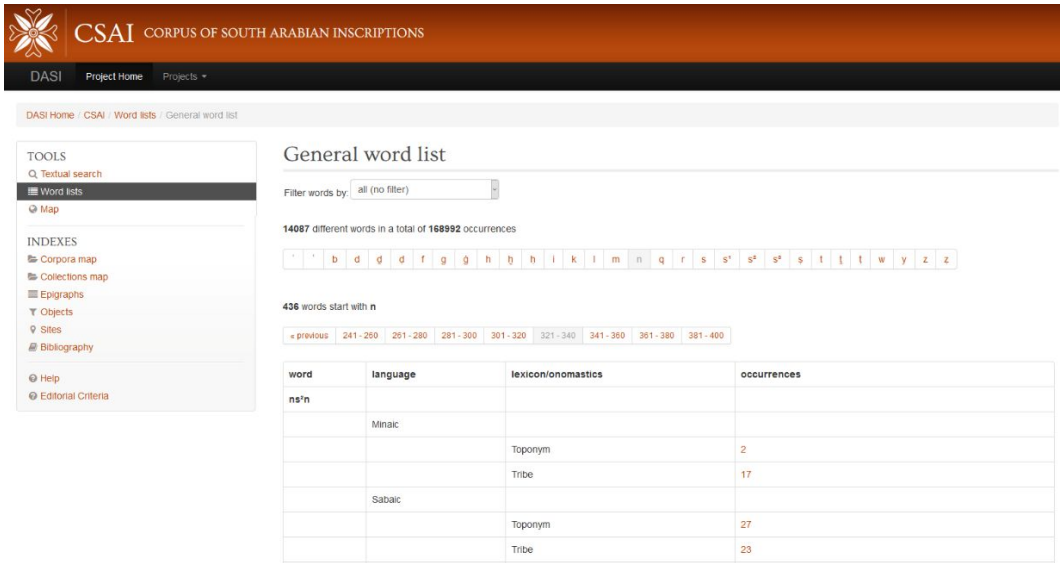

Figure 2. Onomastic item *ns²n* in the Word List of DASI

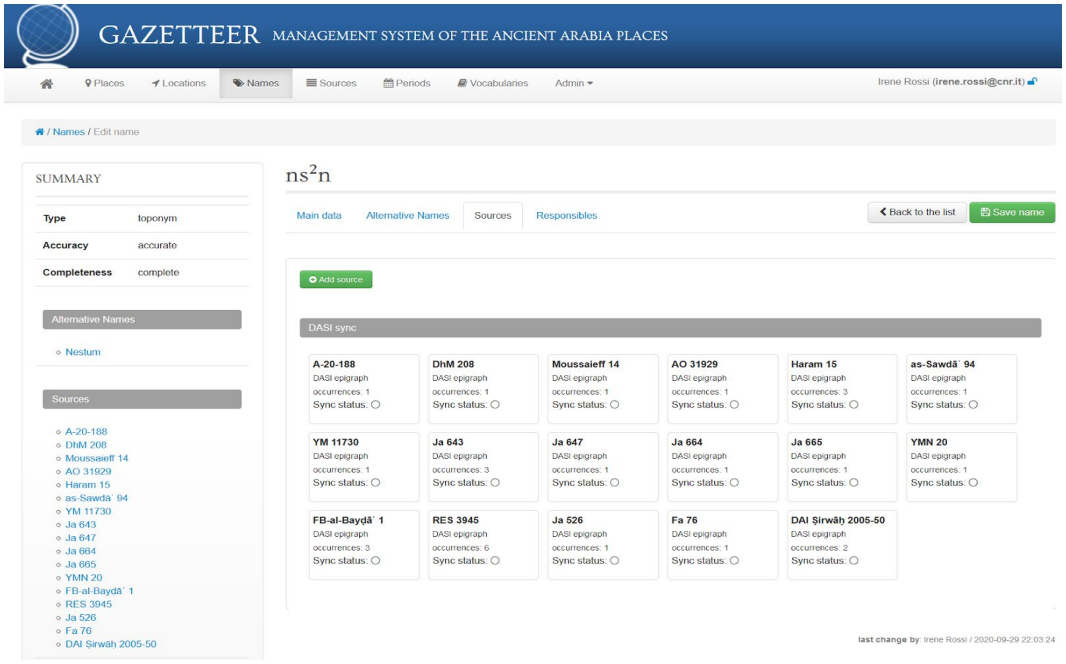

Figure 3. *ns²n* Name record in the Gazetteer of Ancient Arabia, with related Source records

Both the Location and the Name records are connected to the Place record as-Sawdāʾ / Nashshān (archaeological site / ancient settlement). This can be implemented with a proper description and, in turn, can be related to the monuments located in its territory and to the near city of al-Bayḍāʾ / Nashq (Fig. 4).

Since *ns²n* is also the appellation of a tribe (<orgName type="tribe"></orgName>; Fig. 2), further Name and Place records are created for this tribe; the Place record can be linked firstly to as-Sawdāʾ / Nashshān (archaeological site / ancient settlement), and to further ancient settlements it was connected with.

5. 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 word, is the direct bond with the annotated epigraphic corpus of DASI. The core data of the Gazetteer result automatically from the mass digitization of the direct written heritage of pre-Islamic Arabia conducted during the project DASI, according to guidelines that have established themselves as proper "standards" in the digital epigraphy field, and applying the best practices that were later formalized under the label of FAIR principles [7]. 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.

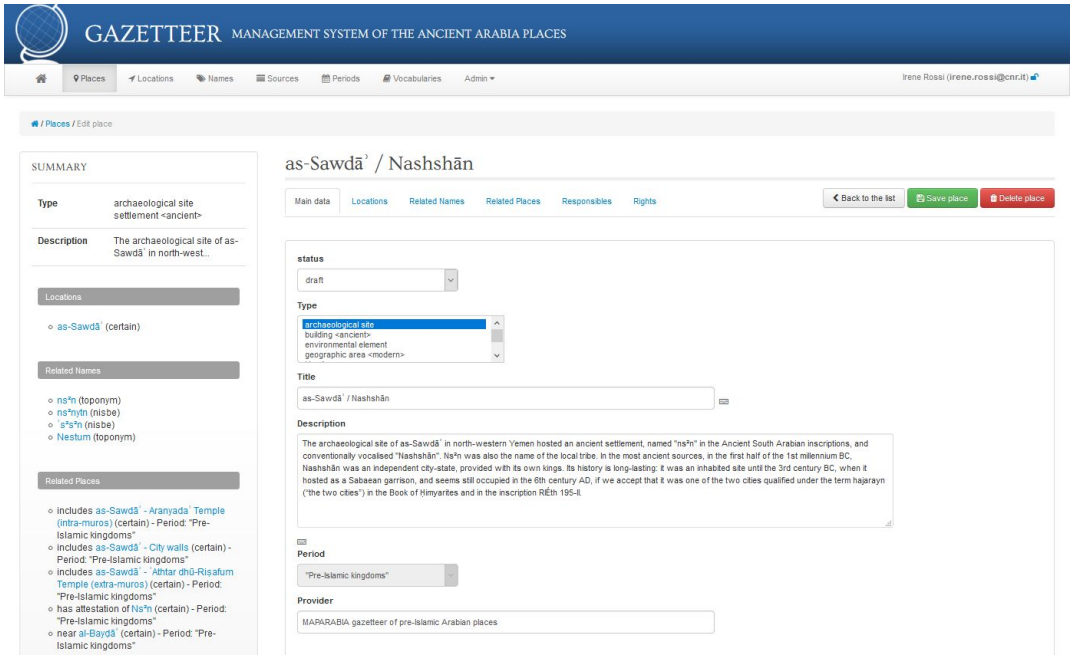

Figure 4. Place record as-Sawdāʾ / Nashshān in the Gazetteer of Ancient Arabia

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.

BIBLIOGRAPHY

[1] Avanzini, Alessandra, Annamaria De Santis, and Irene Rossi. ʻEncoding, 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. Berlin, Boston: De Gruyter, 2018. https://doi.org/10.1515/9783110607208.

[2] Schiettecatte, Jérémie. *D'Aden à Zafar. Villes de l'Arabie du Sud préislamique*. Paris: De Boccard, 2011.

[3] Avanzini, Alessandra. *By Land and by Sea: A History of South Arabia before Islam Recounted from Inscriptions*. Roma: «L'Erma» di Bretschneider, 2016.

[4] Elliott, Tom, Gabriel Bodard, Elli Mylonas, Simona Stoyanova, Charlotte Tupman, Scott Vanderbilt, *et al.* 'EpiDoc Guidelines: Epigraphic Documents in TEI XML (Version 8)', 2007-2017. http://www.stoa.org/epidoc/gl/latest/.

[5] Southall, Humphrey, Ruth Mostern, and Merrick Lex Berman. 'On Historical Gazetteers'. *International Journal of Humanities and Arts Computing* 5, no. 2 (1 October 2011): 127–45. https://doi.org/10.3366/ijhac.2011.0028.

[6] Shaw, Ryan. 'Gazetteers Enriched: A Conceptual Basis for Linking Gazetteers with Other Kinds of Information'. In *Placing Names: Enriching and Integrating Gazetteers*, by Merrick Lex Berman, Ruth Mostern, and Humphrey Southall, 51–63. Bloomington: Indiana University Press, 2016. https://doi.org/10.2307/j.ctt2005zq7.

[7] Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, *et al.* 'The FAIR Guiding Principles for Scientific Data Management and Stewardship'. *Scientific Data* 3, no. 1 (15 March 2016): 160018. https://doi.org/10.1038/sdata.2016.18.