Digital Edition and Database: Note on the *Keshif* Database

Aysu Akcan

This note outlines the editorial and encoding principles developed over the past year and used to produce the digital editions for the *Keshif* Database, which are based on micro editions published in *Keshif: E-Journal for Ottoman-Turkish Micro Editions*.The *Keshif* Database runs in a WordPress environment hosted by the University of Vienna and parses TEI P5–encoded XML files to generate predefined filters, allowing users to search and filter documents directly from the XML corpus.

The note is organized in three parts. The first part describes the <teiHeader>, which carries the bibliographic, institutional, and descriptive metadata for each file and generates the database's *Metadata* display. The second part describes the <body>, where the original text and its English translation are marked up with the entity and concept layers—persons, roles, dates, places, genre and format designations—that populate the remaining filters and make the corpus searchable. The third part, "A Note on the Note," addresses the status of the decisions recorded here and the wider set of questions they open for future work.

The <teiHeader>: Metadata

The <teiHeader> provides the basic metadata framework of each TEI XML file. It identifies the digital file, records its source, and defines the main descriptive categories used for database integration. Elements such as <fileDesc>, <encodingDesc> and <profileDesc> serve this function briefly by recording the digital title, editorial responsibility, and publication details of the file.

The main metadata work, however, is concentrated in the <sourceDesc> element located within <fileDesc>, particularly in the more specialized <msDesc> (manuscript description). In keeping with the TEI-Guidelines, this section records the manuscript or archival witness through elements such as <msIdentifier> and <msContents>. It provides the basis for metadata fields such as repository, identifiers, and source related

attribution. As the *Keshif* Database develops, this part of the header will be further regularized through a more systematic structure aligned with cataloguing practice.

A second key layer is <profileDesc>, particularly <langUsage> and <textClass>. Within <langUsage>, the languages represented in the file are declared explicitly, and the language of the original text is also tagged separately and displayed in the metadata. This is important because the corpus points not only to the Ottoman *elsine-i s̲elās̲e* framework, but also to a wider multilingual textual world that includes texts written entirely in Arabic or Persian, as well as materials in Chagatai, Hebrew, and other languages.

Within <textClass>, genre and format terms are distinguished from other descriptive keywords in order to support filtering in *Keshif*. This distinction will be examined in detail in the section below.

The <body>: Structure and Markup

Within the <body>, two elements are used. The first division, , contains the original text; the second, , contains the English translation. The *Persons, Roles, Dates, Places, Genre and Format*, and related filters in the database's left-hand menu are generated primarily from the content of .

The markup strategy prioritizes entity and concept layers that can be extracted consistently across heterogeneous materials. Persons are encoded with <persName>, offices and ranks with <roleName>, and honorifics, epithets, and other additional name components with <addName>. The distinction requires careful judgment. TEI defines <roleName> as a title or rank that exists, at least to some degree, independently of its bearer, while <addName> covers honorifics, epithets, aliases, and other descriptive elements within a personal name. The TEI-Guidelines also note that this boundary is not always clear and may require culture specific decisions. Terms such as *aġa, aʿyān, şeyḫ,* and *ʿālim* illustrate this difficulty. In Ottoman-Turkish, such elements may function either as <addName> or as <roleName>, depending on context, rather than denoting discrete offices in a fixed sense. In the *Keshif* Database, however,

they are encoded not only as <addName> but also, depending on context, as <roleName>. This does not resolve their broader semantic range but merely establishes a limited and consistent practice for the database. By contrast, terms such as *müderris, ḳāḍī, defterdār, kātib/kātip*, and *beg/beġ* are encoded as <roleName> when they clearly denote an office or rank identifiable beyond the individual bearer.

Dates are encoded only with <date>. Hijri dates are marked with @calendar="#hijri" and appear in the database with a green “h”, while Gregorian dates appear with a green “m”. Where only a Hijri date is preserved in the source, approximate Gregorian correspondences may be added later through @notBefore and @notAfter. These values indicate heuristic chronological placement, not exact conversion. This approach preserves the source form while improving searchability and comparison across the database.

The database heading *Places* serves as a deliberately broad category for geographic references. For this reason, <placeName> is used rather than TEI’s more specific place-name elements. This decision is both technical and methodological. In the TEI-Guidelines, <placeName> functions as the general element for a place name, whereas <settlement>, <region>, and <country> denote narrower categories. Because the corpus includes towns, cities, provinces, countries, and Ottoman territorial designations whose status does not always remain stable across texts, these finer distinctions are not imposed at the point of initial tagging. Instead, <placeName> is retained as a broad and consistent category, while @type is used to preserve historically specific designations such as *eyālet* or *ḳarye* where necessary. This does not reject finer differentiation in principle; rather, it defers it until a more robust typology can be developed within the database.

The *Genre and Format* filter, and the associated category in the database, draw on both the header metadata and the body text. The primary aim is to work from within each text rather than to impose a fixed external typology. Each Ottoman text carries its own structural and textual information, inseparable from its format. A *kitābe*, a *ġazel*, or a *mektūb* is not merely an instance of a genre. Each text actively produces its generic identity through formulaic openings, organizational conventions, and material layout.

In encoding these texts, it is therefore important not only to assign terminology from the outside but also to retain the terminological layers that the text itself produces in relation to its format.

This matters especially for early modern Ottoman texts, where genre boundaries are often fluid, where a single document may bear multiple overlapping textual identities, and where the vocabulary used to describe format is historically specific and analytically significant. Allowing each text to surface its own textual and structural format designations, rather than absorbing them into a predetermined classification, is an attempt to keep that historical specificity visible within the database.

To support this, body-level format information is encoded with the <term> element. Among its available attribute options, type="textGenre" is used exclusively. This is narrower than the full semantic range that @type could accommodate in principle, since the attribute may carry values such as type="register", type="topic", or type="domain", as well as project-specific categories. Retaining type="textGenre" nonetheless reflects both a conceptual and a practical consideration.

In the early modern Ottoman textual world, format and genre are often inseparable: arrangement, formulaic structure, and textual presentation frequently help constitute genre itself. For this reason, and to maintain a consistent degree of standardization across the corpus, body-level format designations are subsumed under the broader category of text genre. The aim is not to exhaust the classificatory potential of @type, but to use it selectively in support of a text-centered representation of early modern texts and their structural and textual properties.

Micro-editions frequently involve uncertain readings, variant spellings, and editorial normalization. TEI's editorial mechanisms are treated as core transcriptional infrastructure rather than optional annotation. Alternative editorial encodings are represented with <choice>, using <sic>/<corr> for apparent error and correction, and <orig>/<reg> for original and regularized forms. Where the text is only partly legible or cannot be transcribed with certainty, <unclear> is used; where no reading can be recovered, <gap> is used. Material supplied by the editor is marked with <supplied>, while additions, deletions, and substitutions in the source itself may be encoded with <add>,

<del>, and <subst> as needed. These practices preserve scholarly accountability while enabling stable indexing and cross-file comparison. As the project develops, additional encoding of this kind will be reflected in the database's left-hand menu.

A Note on the Note

Because the *Keshif* Database aims not to answer a single research question but to make micro-editions searchable and filterable according to a set of basic standards, it necessarily entails choices that are open to discussion. The fact that these micro-editions arrive in *Keshif* from highly varied sources and are then re-encoded for database use has raised numerous questions about preparing digital editions of Ottoman texts. Issues concerning encoding decisions, taxonomic choices, the limits of standardization, and the relationship between textual representation and scholarly accountability will be addressed in a more comprehensive study now in preparation, where matters left deliberately undiscussed in this note will be treated in fuller detail. The most important point to register here is that feedback from both article authors and users with expertise in Ottoman studies is essential to the development of this database. The *Keshif* Database is built in full awareness that its texts belong to a much larger textual field. The value of attending carefully to the formal properties of texts and—perhaps most importantly—of establishing shared tagging conventions and a common encoding language capable of representing those properties accurately is evident. It is a goal this project continues to pursue.