Development of a normalized hadith narrator encyclopedia with TEI

Hajer Maraoui[1], Kais Haddar[2], Laurent Romary[3]

[1] Faculty of Sciences of Tunis, University of Tunis El Manar, MIRACL Laboratory, Tunisia
hajer.maraoui@fst.utm.tn
[2] Faculty of Science of Sfax, University of Sfax, MIRACL Laboratory, Tunisia
kais.haddar@yahoo.fr
[3] Inria, Team ALMAnaCH, Germany
laurent.romary@inria.fr

Abstract. The investigation in the narrator list of prophetic tradition (hadith) is considered an important task for different hadith sciences such as the biography of narrators. In fact, the authenticity of hadith is intensely related to the chain of narrators how transmitted the story. In addition, this science is interested essentially in analyzing the hadith narrator profile. Indeed, having a standardized encyclopedia of hadith narrators can help researchers to explore and manipulate the various hadith documents and can simplify different analyzes. For this reason, we aim to develop a standardized hadith narrator encyclopedia with Text Encoding Initiative (TEI) language. To achieve this, we propose a TEI model for the hadith narrator properties. To experiment our TEI model, firstly, we construct a corpus of articles about hadith narrators from Wikipedia. Secondly, we use a system allowing the named entity recognition in relation with narrator data. Thirdly, we perform a post-processing to complete the corpus TEI annotation. Fourthly, we generate the hadith narrator encyclopedia based on our TEI model. The obtained result are encouraging despite some problems related to exceptional cases.

Keywords: Hadith narrator encyclopedia, narrator data standardization, TEI model, ANER system, Wikipedia database.

1 Introduction

To build knowledge used by Hadith sciences, the researchers collect the required information from different resources. The Internet is on the top of these resources via Islamic websites and free encyclopedia. The majority of questions under the theme of hadith are directed at its components Isnad or Sanad (the chain of narrators) and Matn (the narration text). Besides, more interrogations are focusing on the chain of narrators of the hadith text, cited in the Isnad. From the hadith science perception, examining the chain of narrators is considered a main task for different disciplines such as the science of the biography of narrators (*al-jarḥ wa al-taʻdīl*, discrediting and accrediting) which proceed a number of aspects like hadith authentication and classification [1]. It is also substantial for the individual searcher who looks for the information related with hadith transmitters from the first generation of narrators: the “Sahaba” (or companions) of the Prophet Muhammed (peace upon him) to their descendants. This data can be extended to cover more details concerning the narrator properties and relationships. In fact, this data needs to be collected in one standardized database to support the interoperability of the information.

From this perspective, we aim to create a normalized encyclopedia containing all the relative information of the hadith narrators, in order to prepare it to be used by different Islamic disciplines. We use the Text Encoding Initiative (TEI guidelines) [2] to achieve a standardized formalism of the data. Besides, we use the free resources of Wikipedia to collect the required information about the hadith narrators. We follow a symbolic approach method to realize the hadith narrator encyclopedia. This method starts with a TEI model development in order to encode the hadith narrator data. Then, we construct a corpus including hadith narrator properties. To create the hadith narrator encyclopedia, we annotate in TEI the constructed corpus using a named entity extraction tool [3]. Finally, we develop a prototype to generate the normalized narrator encyclopedia. The developed prototype experiment the proposed TEI model.

This paper is composed of six sections. Section 2 presents an overview on the related works. Section 3 gives details about our proposed TEI model for the data encoding of hadith narrator. Then, section 4 introduces our system for the construction of a normalized hadith encyclopedia. Section 5 presents the system evaluation step. We cloture our paper with a conclusion and perspectives in section 6.

2 Related works

The term hadith is an Arabic word means report. Hadith mentions the speech or the action of the Prophet Muhammad (peace upon him) transmitted first orally between his companions and among the next generations of hadith narrators [4]. After that, the scientists of hadith traveled and collected hadiths to write it down in corpora [4]. The well-known corpus, Sahih al-Bukhari, is one of the six major hadith collections of Sunni Islam. Muhammad al-Bukhari, the author of this corpus, had collected 600000 hadiths of which he only considered 7275 ones as authentic in his work [5]. In next subsections, we present a summary on the Isnad of the hadith and some related works.

2.1 Overview on hadith Isnad

The hadith text is characterized with two main components: the Matn /المتن/ and the Isnad /إسناد/ (or Sanad /سند/). The Matn contains the text of the narration and the Isnad states the list of narrators who transmitted the hadith. The Isnad supports the hadith authenticity. In fact, the hadith sciences proved that Isnad is essential to verify whether a hadith is sound or not. The authentication method follows a careful examination of the chain of transmission to find out if the temporal and spatial links between the narrators are possible and to judge the reliability of the reporter. Which also means that any weakness in the Isnad conclude that the relative Matn is rejected [1][4-6]. Hadith processing was the topic of many projects focused on several fields of researches (hadith ontology, linguistic analyzing, etc.). The following subsection presents some related works that experiment the Isnad treatment.

2.2 Related works

Many researchers used several NLP methods and techniques to create different software and websites that helps to determine hadith judging and classification. For this field of research, we mention the following examples:

The research that carried out by [7] developed a prototype to build database to encode hadith and narrators with XML format. This work is based on a database that manages the hadiths of “Sahih Al-Boukari” book and the narrator information extracted from “Tahdheeb Altahdeeb” book. This prototype identified the chain of narrators and treated the collected information with HPSG formalism. The prototype offered a graphical user interface allowing the access and the visualization of the list of narrators and hadiths from the treated books.

The study conducted by [8] proposed a system allowing the recognition of hadith narrator chain and their classification. This system presented the extracted sequences of the narrator names in a graphical format as a network. The edges in this graph presented the transmission links between narrators.

The study performed by [9] putted forward an Isnad ontology to support the hadith authentication. The authors extended the terms of the ontology to cover more semantic relations and properties of hadith narrators to process the Isnad judging. The authors experimented an evaluation step with DL-Queries and hadith text examples.

Research that performed by [10] proposed an annotated narrator graph extractor (ANGE) from hadith text and biography books. This technique is based on different NLP technologies like graph algorithms, cross-document reconciliation, finite state machines and morphological features.

According to the different works focused on the Isnad part, the chain of narrators is a key part that needs to be extracted and annotated for the authentication process of the hadith. However, only few studies, frequently in ontology, extended the narrators presentation to extract their properties and the semantic relations between them. Yet, in the one hand, according to our knowledge, no study has concentrated on applying a standardization formalism like TEI to normalize the extracted information. In the other hand, there is more details in the person properties and the relationships that can be mentioned to define the narrator historical profile.

3 TEI modelling of hadith narrator data

One of the benefits of applying TEI normalization on any type of data is the flexibility of TEI modelling. TEI covers large data categories and classes, and allows the restructuring of the elements without losing the specification of each component [11]. We took advantage from these features to propose a TEI encoding model formed for hadith narrator. For the final encyclopedia, this model represents the primer unit encoding which refers to a hadith reporter. Therefore, to create the required TEI model, we start with the selection of the elements from the core module that coordinates with the person data encoding. Then, we form the structural design of the final model. This section contains, in a first part, an overview on the collection of the elements dedicated for the named entities and person properties description. Then, in a second part, we present our proposed TEI model.

3.1 Encoding the person properties with TEI

TEI guidelines [2] recommend a full module to encode the named entities in the text [12]. Person name is one of the most detailed entities in this module. Moreover, TEI presented different elements to encode other types of named entities that indicate an information about that person such as birth date, place of residence, etc. Indeed, the main TEI element to represent all information related to a person is <person> [12]. Table 1 illuminates this element and some attributes connected with it.

Table 1. Summery table of the person head element

Element	Description	
<listPerson>	(List of persons) represents a list or group of identifiable individuals where each one is marked up in a <person> element.	
<Person>	Includes all the information about a definite person.	
	Attribute	Description
	xml:id	Specifies one unique identifier for the <person> element.
	role	Mentions an additional information about the person like his occupation, or his social rank.
	sex	Indicates the gender of the person.
	age	Indicates the age of the person.

The <person> element is the main component to define a person. Besides, the <listPerson> element can be used to enumerate a list of persons as precise index or bibliography. The <person> element is characterized with a number of attributes that can serve to present some personal information like “xml:id” for the personal identifier, or “age”, “sex” and “role” for farther specifications. The rest of the data related to the person can be observed as two categories: personal and social information. Table 2 summarizes TEI elements for the encoding of the personal information.

Table 2. Summery table of the TEI elements for the personal information

Category	Element	Description
Name phrase of the person	<persName>	(Personal name) contains a part or full name of a person.
	<surname>	Contains the family name of the mentioned person.
	<forename>	Contains a forename of the mentioned person.

		Attribute: *type*	Describes the type of the forename using a significant terms.
		Attribute: *Sort*	Identifies the order of the forename comparing to others forenames.
	<roleName>	Contains the society rank or official title of the cited person.	
	<addName>	(Additional name) contains an additional name for the person.	
	<nameLink>	(Name link) contains a connecting phrase or link occurs in the phrase of the person name such as "*de*" or "*بن*".	
Dates related to the person	<birth>	Contains the birth details of the person like the date and place.	
		Attribute: *when*	Redefines the birth date in a standard form
	<death>	It is similar to birth element but contains the death details.	
		Attribute: *when*	Redefines the death date in a standard form
	<floruit>	Define the period during which a person lived.	
		Attribute: *notBefore*	Represents the initial date for the lived period in a standard form.
		Attribute: *notAfter*	Represents the ending date for the lived period in a standard form.
	<date>	Represent a date value in any format.	
		This element can also have the attributes: *when* (defined above).	

The <persName> element can distinguish a name phrase of a person in the text. It is possible to include one or more name component (like forename, surname, added name, etc.) in the corresponding elements imbricated inside <persName> element. It is also the case for the <birth> and the <death> elements which contain all the information about the dates in <date> elements and their places in <placeName> elements (defined bellow in Table 3). Moreover, these elements can have the "when" attribute to present a standard form of the date value. The second category of data is the social information. Table 3 summarizes TEI elements for this category.

Table 3. Summery table of the TEI elements for the social information

Category	Element	Description	
Language competences of the person	<langKnowledge>	(Language knowledge) presents the linguistic knowledge of the person in <langKnow> elements.	
	<langKnow>	(Language known) indicates one linguistic competence	
		Attribute: *tag*	Provides a practical tag for the relative language.
		Attribute: *level*	Precise the level of knowledge of the language
Places related to the person	<residence>	Refer to the places of residence of the person.	
	<placeName>	Contains an absolute or relative place name.	

		This element can have the attributes: *notBefore and notAfter* (see Table 2)	
	<country>	Specifies the geo-political unit, such as the nation, country, commonwealth, etc.	
		Attribute: *Key*	Identify a meaningful value defined externally to identify the named entity.
	<settlement>	Contains a single geo-political or administrative unit, such as a city, town, or village.	
		This element can have the attribute: *type* (see table 2).	
Social identity of the person	<nationality>	Describe of a person nationality or citizenship.	
	<education>	Describe of the educational experience.	
	<affiliation>	Present the person affiliation.	
	<faith>	Specifies the faith, religion of a person.	

TEI provides several elements to markup different information about the language skills, the residence and the social identity of a person. Additionally, information about relationships between the people makes also a part in the TEI encoding for the person data. Table 4 summarizes the two main TEI elements for this information.

Table 4. Summery table for the TEI elements for the relationships between persons

TEI element	Description	
<listRelation>	Provides information about relationships either identified amongst people, places, and organizations, informally as prose or as formally expressed relation links.	
<relation>	(Relationship) describes any kind of relationship or linkage amongst a specified group of places, events, persons, objects or other items.	
	Attribute	Description
	type	Provides a unique identifier for the element bearing the attribute
	name	Supplies a name for the kind of relationship of which this is an instance.
	active	Identifies the 'active' participants in a non-mutual relationship, or all the participants in a mutual one.
	passive	Identifies the 'passive' participants in a non-mutual relationship.
	mutual	Supplies a list of participants amongst all of whom the relationship holds equally.
	cert	(Certainty) signifies the degree of certainty associated with the object pointed to by the certainty element.

The element <listRelation> englobes a sequence of <relation> element, to express relation links between mentioned persons. In fact, the <relation> element performs the main unit of the relationship encoding. It can come with the attributes "type" and "name" to specify the classification and the designation of the relationship. As well, it can contain the attributes “active”, “passive” and “mutual” to specify whether the links are conjoint or not. Here, only one of the attributes “active” and “mutual” can be supplied. The attributes “passive” and “active” can only supplied together. This constraint is not enforced for all schema language. For example, the following personal relation illustrates the link between the father “UrwaBinAz-Zubair” and his son “HishamBinUrwa”:

```
<relation type="personal" name="FatherOf" active = "#Ur-
waBinAz-Zubair" passive="#HishamBinUrwa" cert="high"/>
```

To achieve the TEI modelling of the hadith narrator data, we select the adaptable TEI elements from TEI module for encoding persons, dates, places and personal relationships [11]. The proposed TEI model is presented in the next section.

3.2 Proposed TEI model for the hadith narrator data

For the representation of the hadith narrators in an encyclopedic structure, we start with the basic elements that include the unit sequence. Then, each component is presented in the element that englobes all the relative information. Fig. 1 illustrates the basic elements model that we adapt with the data provided for the hadith reporter in the Islamic resources and Arabic literature.

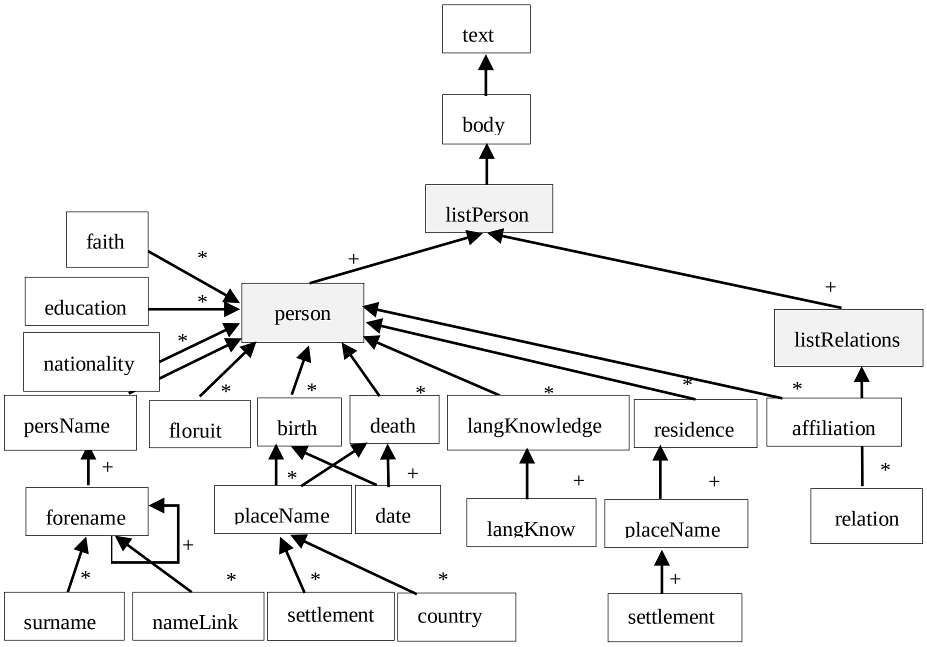

Fig. 1. Extraction from TEI encoding model for the details of hadith narrator

The representation begins by encoding the properties of the encyclopedia in the <teiHeader> element. Then, it uses the <body> element including in the <text> element. The <body> element contains the main <listPerson> component that covers the entire list of narrators and its information. This element contains a sequence of <person> element and a <listRelation> element that represents respectively the information about hadith narrators and the relationships between them. The <person> element has an "xml:id" attribute to assign a unique identifier for each narrator, which will then be used for narrator identification in the relationship list. Moreover, this identifier will be the connector link used in the Isnad encoding of each hadith.

In addition, the <person> element includes all the personal and the social properties of the hadith narrator. The personal properties such as the full name, the birth date, the death date and their places are encoded respectively in <persName>, <birth> and <death> elements and the corresponding place names in <placeName> element. The <floruit> element contains the period that the narrator lived according to the birth and the death dates. The social properties such as the nationality, the residence, the education, the religion, the affiliation and the spoken languages are described respectively in <nationality>, <residence>, <education>, <faith>, <affiliation> and <langKnowledge> elements.

4 Elaborated method for the construction of a hadith narrator encyclopedia

To achieve the final hadith narrator encyclopedia, we start with a conceptual study of a creation system for the encyclopedia. The realization of this system follow a symbolic approach. It based on a method composed of three main phases. Fig. 2 illustrates the general architecture of the system.

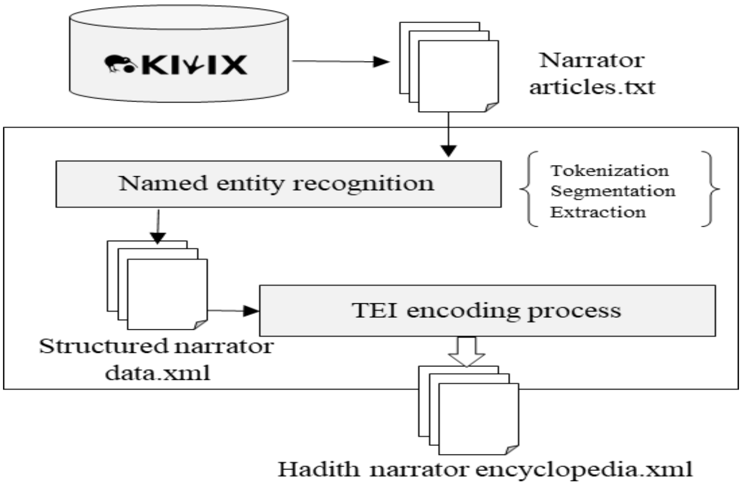

Fig. 2. Developed method for the construction of hadith narrator encyclopedia encoded in TEI

The method allowing the construction of the encyclopedia is composed of three successive stages. In this method, the first step starts with a connection to the Wikipedia resources to provide all the data relative to the narrator profile. This procedure is the step of collecting the articles about the hadith narrators. We rely on this database for the reason that it provides all the authentic information about hadith narrators. In addition, Wikipedia offers the free accessibility to the required information. Also, through its Kiwix platform, it allows the backup of resources as .txt files. For this reason, we use the open source version of the Wikipedia database for Arabic language.

The second phase performs a named entity recognition step. This step is based on a system proposed by [3]. This system generates the extracted named entities in XML output and borders each detected component with TEI encoding. In addition, we perform a post-processing to extend the data recognition process for the rest of the social information and the relationships. The obtained XML files are used as an input of the next phase.

The third step is for the TEI encoding process of the collected narrator properties. In this step, we developed a tool for the automatic encoding of the data. This tool applies the TEI encoding to generate the base of the narrator encyclopedia. This encoding procedure follows our specific TEI model for hadith narrator properties.

5 Experimentation and evaluation

For the implementation, we use some tools and programs. Indeed, for the manipulation of the TEI files, we use Oxygen XML Editor. Besides, we develop the prototype using JAVA language and the JDOM Library.

The output of our system is the encyclopedia of narrators encoded in TEI. The following TEI code illustrates an example of generated data encoding for the hadith narrator: “الزبير بْنِ العوام” (Az-ZubairBinAl-Awwam).

```
<person xml:id="Az-ZubairBinAl-Awwam" role="hadithNarrator"
sex="1" age = "Adult">
   <persName xml:lang="ara">
      <forename>الزبير</forename>
      <forename sort="1" type = "nasab">
          <nameLink>بْنِ</nameLink>
          <forename>العوام</forename></forename>
      <forename sort="1" type = "nisba">الأسدي</forename>
      <surname>القريشي</surname></persName>
   <birth when="0594">
      <date when="0594">594 AD - 28 BH</date>
      <placeName>
          <settlement type="city">Makkah</settlement>
          <country key="KSA">The Arabian Peninsula</country>
      </placeName></birth>
   <death when="0656">
      <date when="0656">656 AD - 38 AH</date>
      <placeName><settlement type="city">Bassorah</settlement>
         <country key="Ir">Iraq</country>
      </placeName></death>
   <faith>Islam</faith>
   <langKnowledge>
      <langKnown tag="ar" level="H"> Arabic</langKnown>
   </langKnowledge>
   <nationality>Arabian Peninsula</nationality>
   <residence notBefore="0584" notAfter="0656">
      <placeName><settlement type="city">Makkah</settlement>
      </placeName></residence>
   <affiliation>Politician</affiliation>
   <education>Islamic legislation</education>
   <floruit notBefore="0584" notAfter="0656"/></person>
```

After the sequence of <person> elements, the system encodes all the relationships detected between the narrators in the <listRelations> element. These relationships refer to all personal and social connections between different people (such as "father of", "mother of", "son of", "uncle of", etc.). The following TEI code presents an extraction of the generated relationships encoded in the <relation> sub-elements.

```
<listRelation>
<relation type="personal" name="SpouseOf" mutual="#Az-
ZubairBinAl-Awwam #AsmaBintAbuBakr" cert="high"/>
<relation type="personal" name="FatherOf" active="#Az-
ZubairBinAl-Awwam" passive="#UrwaBinAz-Zubair" cert="high"/>
<relation type="personal" name="MotherOf" active = "#AsmaBintA-
buBakr" passive="UrwaBinAz-Zubair" cert="high"/>
<relation type="personal" name="SonOf" active="#UrwaBinAz-
Zubair" passive="#Az-ZubairBinAl-Awwam #AsmaBintAbuBakr"
cert="high"/>
<relation type="personal" name="UncleOf" active="#AbdullahBinAz-
Zubair" passive="#HishamBinUrwa" cert="high"/> …</listRelation>
```

To evaluate our system, we use Kiwix platform of Wikipedia to collect the articles about the hadith narrators from the list of the prophet companions. This database provides us with information about 642 hadith narrators. Therefore, we obtain an article for each narrator. As a result, the system generates the hadith narrator encyclopedia encoded with TEI, representing all the collected data. Table 5 illustrates the obtained results.

Table 5. System obtained result.

Hadith narrator Information	Encoded correctly	Encoded incorrectly
642 article of hadith narrators from Wikipedia	96%	4%

The system succeeded to generate a total correct encoding of the information of 616 hadith narrators while respecting the TEI model. This number is equal to 96% of the treated data. However, 4% of the information is not detected or incorrectly encoded. This is related with some particular forms of the information inside the hadith narrator articles, or a wrong detection or encoding of the information related with some ambiguous parts. Regarding the results, the coverage of the information in the hadith narrator encyclopedia is divided on three themes: personal information, social information and relationships. Fig. 3 presents there measurements.

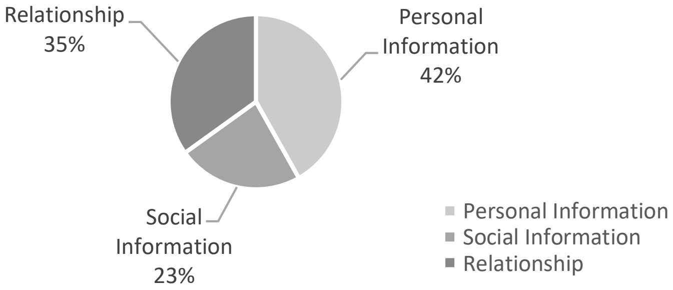

Fig. 3. Information coverage inside the hadith narrator encyclopedia

The majority of information is about the hadith transmitter personal information (full name, birth and death) with 42% of the total information. Then, the relationships cover a 35% of the encoded data. The last 23% of the data is for the social information (affiliation, language, education, etc.). We measure the quality of our work with the values of precision, recall and F-measure presented in Table 6.

Table 6. Summary table of the precision, recall and F-score.

precision	recall	F-measure
0.96	0.96	0.96

The calculation provides an identical precision and recall equal to 0.96. Consequently, F-measure as well is equal to 0.96. Therefore, these measurements prove that the obtained results are encouraging.

6 Conclusion and perspectives

The normalization of the data collection of hadith narrator in one organized encyclopedia can support the manipulation of the information and reduce the difficulties of deep studies. To realize our main objective, we based on TEI standardization. In this work, we started with an overview on the hadith chain of narrators in the Isnad. Then, we continued with a study on the TEI module for encoding person properties. After that, we proposed an adaptable TEI model to encode the information about hadith narrator. Then, we followed a symbolic approach to accomplish the creation of the hadith narrator encyclopedia. The first phase was to collect the data from Arabic Wikipedia. The second phase was to extract the required information including a system of named entity recognition. The final phase was to encode the collection of data with the proposed TEI model. Then, we tested our system with a 642 articles about hadith narrators from Wikipedia. As mentioned, the obtained measures showed that the results were encouraging. This normalized encyclopedia can be used in many other applications.

As perspectives, we want to select more information about hadith narrators from other authentic resources to enrich the data coverage in the encyclopedia. Moreover, we want to expand our TEI model to cover more details related to hadith narrators by integrating other specifications. In addition, we want to improve our system to resolve the problems. Besides that, we want to use the normalized encyclopedia to test a developed system for analyzing hadith text.

References

1. Alrfaai S.: عناية العلماء بالاسناد وعلم الجرح والتّعديل و أثر ذلك في حفظ السّنة النبوية, AlMadina Almonawara, (2004).
2. Burnard L. and Sperberg-McQeen C.M.: TEI P5: Guidelines for Electronic Text Encoding and Interchange, Text Encoding Initiative Consortium, Version 3.0.0. revision 89ba24e, (2016).
3. Ben Mesmia F. et al.: Arabic Named Entity Recognition Process using Transducer Cascade and Arabic Wikipedia, proceedings of Recent Advances in Natural Language Processing, 48-54, Hissar, Bulgaria, Sep 7–9 (2015).
4. Abu Zaho M., الحديث والمحدثون, دار الفكر العربي, Riyadh, KSA, volume 1, (1984).
5. Alaskalani A. : فتح الباري بشرح صحيح البخاري, KSA, volume 1, (2001).
6. Alaskalani A.: تقريب التهذيب, Society AlRisala, Bayrout, Labunan, (2008).
7. Moath N., XML Database for Hadith and Narrators, American Journal of Applied Sciences, 13(1): 55-63, (2016).
8. Muazzam A. at al.: Extraction and Visualization of the Chain of Narrators from Hadiths using Named Entity Recognition and Classification, International Journal of Computational Linguistics Research, volume 5, Issue 1, 14-25, March (2014).
9. Baraka R.: Building Hadith Ontology to Support the Authenticity of Isnad, International Jornal on Islamic Applications in Computer Science and Technology, Vol.2, Issue 1, 25-39, December (2014).
10. Zaraket F. at al.: Hadith Narrator Chain Extraction using Arabic Morphological Analysis, proceedings of the Twenty-Fifth International Florida Artificial Intelligence Research Society Conference. 256-261, (2012).
11. Burnard L. and Sperberg-McQeen C. M.: La TEI Simplifiée : Une introduction au codage des textes électroniques en vue de leur échange, Cahiers GUTenberg, n°24, pp. 23–151, (1996).
12. Dufournau N., Demonet M-L. and Uetani T. : “Manuel d’encodage XML-TEI Renaissance et temps modernes Imprimés-manuscrits,” Version Beta, UMR 6576, (2008).
13. Maraoui H., Haddar K., Romary L., Encoding prototype of Al-Hadith Al-Shareef in TEI, ICALP Conference, Fez, Morocco, Springer, CCIS 782, pp. 1–13, (2017).