Journal of Religion, Media & Digital Culture (JRMDC)

<http://jrmdc.com>

The Arabic Diatessaron Project: Digitalizing, Encoding, Lemmatization¹

Giuliano Lancioni, Roma Tre University

Contact: giuliano.lancioni@uniroma3.it

N. Peter Joosse, University of Oxford and University of Warwick

Contact: faridajoosse@yahoo.com

Keywords: New Testament; Arabic Diatessaron; manuscripts

Abstract:

The Arabic Diatessaron Project (henceforth ADP) is an international research project in Digital Humanities that aims to collect, digitalise and encode all known manuscripts of the Arabic Diatessaron (henceforth AD), a text that has been relatively neglected in scholarly research.

ADP's final goal is to provide a number of tools that can enable scholars to effectively query, compare and investigate all known variants of the text that will be encoded as far as possible in compliance with the Text Encoding Initiative (TEI) guidelines.

The paper addresses a number of issues involved in the process of digitalising manuscripts included in the two existing editions (Ciasca 1888 and Marmardji 1935), adding variants in unedited manuscripts, encoding and lemmatising the text.

Issues involved in the design of the ADP include presentation of variants, choice of the standard text, applicability of TEI guidelines, automatic translation between different encodings, cross-edition concordances and principles of lemmatisation.

About the Authors:

Giuliano Lancioni, Ph.D. (1995) in Linguistics, is a Professor of Arabic language and literature at Roma Tre University, Rome. He has published extensively on the history of the Arabic grammatical tradition, Arabic digital humanities and computational linguistics. He co-edited *The Word in Arabic* (Brill, 2001) and *Dār al-islām / dār al-ḥarb: territories, peoples, identities* (Brill, forthcoming).

N. Peter Joosse (1960) was a fellow-in-residence at the Netherlands Institute for Advanced Studies (NIAS) and a senior research fellow at the universities of Frankfurt and Leiden. In 2010, he joined the Department of Classics & Ancient History of the University of Warwick as a Wellcome Trust researcher to work on the medical history of the Middle East, specifically on the Arabic versions of the Hippocratic Prognostic. In 2013 he joined the current University of Oxford project on the medical biographies of the Arab physician Ibn abī Uṣaybi'a. He holds a PhD from the Free University of Amsterdam (summa cum laude).

To Cite This Article:

Lancioni, G., and Joosse, N.P., 2016. The Arabic Diatessaron Project: Digitalisation, Encoding, Lemmatisation. *Journal of Religion, Media and Digital Culture* 5(1), pp. 205-227. Online. Available at: <<http://www.jrmdc.com/journal/issue/view/9>>.

1. Introduction

The Diatessaron is arguably the most important example of the Gospel harmony genre in Late Antiquity and the Middle Ages. Commonly attributed to Tatian (due to an early statement by Eusebius of Caesarea), who was originally a disciple of Justin Martyr (d. ca. 162/167 A.D.), the original text of the Diatessaron is generally assumed to have been written between 160 and 175 A.D. (Marmardji 1935, p. vii-xi; Petersen 1994).²

According to the present state of affairs, the original text of the Diatessaron can be reconstructed by perusing the commentaries—especially the extensive commentary in Syriac by St. Ephrem, which does not quote, however, the entire original text³— and translations. Apart from the

important tradition of Western Diatessaronic translations, which has a very complex textual history and gave rise to much discussion about a possible Latin archetype, the earliest complete witness of the Eastern tradition is an unabridged Arabic translation attributed to Ibn al-Ṭayyib.

The importance of the AD and the very unsatisfactory state of its editions and translations prompted a group of researchers to devise an international project to collect, digitally edit and make available to scholars all known manuscripts of the AD, together with translations and other supplementary materials.

The present paper discusses some of the editorial and encoding-related issues involved in the ADP. The text is organized as follows. First (2), a survey of extant manuscripts of the AD is reviewed, including their relation to the current editions of the text. A description of the project follows (3), with specific sections devoted to, respectively, an outline of the general framework (3.1), a discussion of general encoding issues (3.2) and a more specific analysis of how the critical apparatus is encoded according to the TEI specification (3.3). Implementation issues related to the database (4.1) and the diatessaron.org website (4.2) are presented in section (4).

Finally, a full description of known manuscripts of the AD is included as an appendix (5) to this article.

2. The Status Quaestionis of the Manuscripts of the AD

The manuscripts of the AD (Ta) have previously been described in two separate volumes (Joosse 1999; 2012). An electronic copy of the long lost manuscript Q (Coptic Orthodox Patriarchate, Cairo, *muqaddasa* 135; *raqm musalsal* 198) was only made available to us in 2011 through the kindness of the Center for the Preservation of Ancient Religious Texts at Brigham Young University (U.S.A.).⁴ In the same year we were also able to obtain copies of two independent Arabic Gospel harmonies.⁵ Since all known manuscripts of the AD and the other Arabic harmonies are in our possession now,⁶ it has become possible to move into new directions in order to reach a comprehensive understanding of the AD and its place within the history of the Gospels. The eight extant manuscripts of the AD will be described in full for the first time here, so that referencing will be more accessible from now on.

2.1. 1. *MS A, Vatican Arabo XIV*⁷

This manuscript was brought to Rome from the East in 1719 by Joseph Simon Assemani. The MS, from the character of writing and from the presence of certain Coptic letters by the first hand, is supposed to have been written in Egypt. It is usually dated to the twelfth or the beginning of the thirteenth century, although Kahle, after Åkerblad's opinion, assigned it to the thirteenth or fourteenth century. Originally it consisted of 125 folios. In its present state, fols. 17 and 118 are missing, and fols. 1-7 are not well preserved. There are margin notes: emendations, restorations, explanations, some of them by a later hand. The genealogies of Jesus can be found near the beginning of the harmony: the genealogy of Matthew i is in chapter 2 of the Diatessaron, and that of Luke iii in chapter 4. The Evangelists are described by the first two letters of their names. The round dots occurring in the text are red-coloured, and thus are the signs for the Evangelists. The text of the MS is scarcely vocalised and has few diacritical points. The MS does not mention the name of the translator. The folios are numbered by a later hand. There are liquid stains at the bottom corners of nearly every page, which have affected the ink slightly. The *tā' marbūṭa* is seldom written with dots. The *tā'* and *tā'*; *jīm*, *ḥā'* and *ḥā'*; *dāl* and *dāl* are not always distinguished, and coalescence of words frequently occurs. The editors Ciasca and Marmardji used this MS in their respective editions of the AD (1888 and 1935).⁸

2.2. 2. *MS B, Vatican Borgianum Arabo 250*⁹

In 1896, this manuscript was donated by its owner, prominent Catholic Copt Ḥalīm Dūs Ġālī, to the Museum Borgianum de Propaganda Fide in Rome. The manuscript contains the AD on fols. 96b, 97a-353 a, preceded by a long introduction to the Gospels by an anonymous author (fols. 1-95). It consists of 355 leaves: each page is about 9 inches x 6.25 (i.e., 22.5 x 16 cm) and has eleven lines of writing enclosed by gold, blue and red lines connected in the form of rectangles. The big round dots in the text are gold-coloured. The leather binding is claret-coloured and ornamented with golden dots. The MS is usually dated to the fourteenth century. Kahle, however, on the basis of the style of decoration, thought that it could not be older than the sixteenth century. It is most remarkable that the first two pages are written in exactly the same way as *sūra* 1 and the beginning of *sūra* 2 in certain MSS of the Koran. The MS is written elegantly in black *nashī* and resembles the scripture and style of certain sixteenth century Ottoman Koran codices. The genealogies of Jesus can be found at the end of the harmony, as a kind of appendix. The Evangelists are not

indicated by specific signs or letters as was promised in the prologue of the MS. The manuscript is complete. The name of the translator, Abū 'l Faraj Ibn al-Ṭayyib, has been mentioned in the preamble and colophon of the manuscript. The text is fully vocalised. There are, in comparison to MS A, many instances of parablepsis. The folios have been numbered by a later hand. The editors Ciasca and Marmardji used this MS in their respective editions (1888 and 1935).

2.3. 3. MS E, Coptic Orthodox Patriarchate, Cairo (muqaddasa 136; raqm musalsal 131)¹⁰

The manuscript was completed on 27th Bashnes A. Mart. 1511 (i.e., 22 May 1795 A.D.), so it was composed much later than MSS A and B. It was not written very carefully; dozens of instances of sub-standard vocalisation can be found on almost every page. The preamble of the MS is the same as the one in MS B, but it does not exhibit the external form of Koran Mss. The siglum E has been derived from the Dominican Monastery in Jerusalem Saint Étienne, where Marmardji, the discoverer of the manuscript and professor at the École Biblique, lived. The genealogies have been placed at the end of the harmony. The Evangelists are quoted by one characteristic letter. The dimensions of the book are 25 x 18 cm and each page is 19 x 12.5 cm. The total manuscript consists of 114 folios. Marmardji used this MS in his 1935 edition in addition to MSS A and B.

2.4. 4. MS O, Bodleian Library, Oxford, MS Arab e 163¹¹

The MS has been in the possession of the library since 1937. The manuscript contains three Christian texts: an introduction to the Gospels (fols. 5-31), a compendium on the Christian Truth (fols. 41-139) entitled *Tiryāq al- 'uqūl fī 'ilm al-uṣūl*, and finally the AD (fols. 140-288). The copyist Antūnī Sa'd finished the text on the 8 Tobah A. Mart. 1522 (i.e., January 1806 A.D.). At the end of the manuscript he declares that, following the orders he has received, he has made an exact copy of an MS that was completed on the 13 Rajab A.H. 500 (i.e. 15 March 1107 A.D.). This early manuscript had apparently been written by pious members of the Awlād al- 'Assāl, a Coptic family whose prominence flourished in Egypt for several centuries. They were anxious to shape it in such a way that it would make an impression on the Muslims and would thus enhance the value of the AD by associating it with the names of outstanding Christian Arabic scholars. The beginnings of the first and third texts are, for example, written in exactly the same manner as the first and the beginning of the second *sūra* in manuscripts of the Koran. The MS has the same preamble as the MSS B-E-S-(T)-Q. The MS consists of 293 folios, 13 lines on each page, written in

a rectangular box of one blue and two red lines. The codex was supplied with a leather-faced contemporary binding and consists of folia of semi-transparent oriental paper of tough texture and good quality. The genealogies are at the end of the harmony. The Evangelists are quoted by single letters. The folios are numbered by a later hand.

2.5. 5. *MS S, Library Paul Sbath 1020/Salem Ar. 218*¹²

This manuscript was copied by the deacon Ibrāhīm Abū Ṭabal b. Sam‘ān-al- Ḥawānikī, one of the servants of the martyr Merkurios Abū Sayfayn in Old Cairo (al-Fuṣṭāṭ), in A. Mart. 1512 (i.e., A.D. 1797). The siglum S was derived from the first letter of the name Sbath. The MS has the same preamble as MSS B-E-O-(T)-Q. The Evangelists appear to have been described with one characteristic letter. The MS is written in *nashī* script, is bound and consists of 277 pages (142 folios). There are 15 lines on every page. Its dimensions are 168 x 109 mm. The writing surface measures 124 x 85 mm.

2.6. 6. *MS T, Library Paul Sbath 1280/Salem Ar. 446*¹³

The catalogue of Sbath also mentions another manuscript of the Diatessaron: MS 1280. The date of the composition and the name of the copyist are not given. According to Sbath, it must be dated to the eighteenth century. The MS consists of 376 pages (225 folios). The last two pages are missing, breaking off with Mt. 28: 15a (i.e., Ta 53: 30). There are between 12 and 14 lines on each page. The paper is slightly eaten by moths and is in a rather bad state. The MS is written in *nashī* script and is bound. Its dimensions are 150 x 110 mm. The writing surface measures 126 x 75 mm. We chose the siglum T for this MS, the penultimate letter of the name Sbath. This manuscript may be closely related to MS S. Apart from the AD, this MS contains a second text: “a collection of prayers” 32 pages long.

2.7. 7. *MS C, Bibliothèque Orientale, Beirut 429/commonly related to as the “Beirut Fragments”*¹⁴

This manuscript has been preserved in three fragments only, which are commonly called the “Beirut Fragments.” These fragments have been in the possession of the University St.-Joseph in Beirut since 1897. They consist of three folios from a manuscript that was finished in July 1332 A.D. by Abū ‘l-Barakāt Ibn Abī ‘l-[Kabar?]. The folios contain the narrative of the Lord’s Supper

and the last sentence of the Diatessaron. They contain also an interesting colophon. The fragments show a form of the text that generally agrees with that of Codex A. The Evangelists are marked with the first two letters of their names, as in MS A. The manuscript to which the fragments belonged was connected through three manuscripts – copied from one another in Egypt – with a “very old” MS, written in the city of God (Antioch). The oldest of these three MSS, the one that had been copied directly from the “very old” MS, was written by Anbā Yūsāb Ibn al-Muḥabrik, Bishop of Fuwwah (on the Rosetta Nile) in the first half of the thirteenth century. The three folia are written on tough paper, which resembles parchment; the ink is excellent; the title and the names of the Evangelists are presented in red; the pages measure 20 x 11 cm. We mention this manuscript here because it stands in evidence of the existence of a separate tradition, which is being transmitted in this text and likewise in MS A.

2.8. 8. *MS Q, Coptic Orthodox Patriarchate, Cairo (muqaddasa 135; raqm musalsal 198)*¹⁵

In 2011, we acquired a film of the long lost manuscript of the AD: MS Q. It was microfilmed by Brigham Young University [BYU] at the Coptic Orthodox Patriarchate [COP] in al-Azbakīyah, Cairo in 1984. The siglum Q stands for the Arabic words for Cairo and Coptic: *al-Qāhira* and *qibṭī*. The manuscript has been recorded in the catalogue of Simaika-Yassa and the *Edizione Anastatica* 1959 of Georg Graf's *GCAL*.¹⁶ Both handbooks affirm that the manuscript is undated and bears the title *al-Bustān al-zāhir wa 'l-injīl al-ṭāhir al-mad'ūww Diyātāsārūn*, “The flowering garden and the pure Gospel, which is called Diatessaron,” by the Greek Tatian. If we take a look at the preamble of MS B (Vatican Borgianum Arabo 250), we will find the citation, “We begin with the help of God the Sublime, with the writing down of the pure Gospel and the flowering garden which is called Diatessaron: the explanation of this expression is the fourfold and it is this which Tatian the Greek has assembled from the four Evangelists [...]”. Evangelists are quoted by one characteristic letter, and the genealogies of Christ are placed at the end of the text, as a sort of appendix. These traits are altogether absent from the text as found in the manuscripts of the A-family [or: Antioch Group]. The manuscript has been dated by BYU to the eighteenth century, and it appears that it is closely related to MS E.

In his *al-Fihris (The Catalogue)*, Paul Sbath (1938, p. I/23-24) mentions seven more manuscripts of the AD that he had seen but was not able to acquire. One MS remains unidentified, two are apparently the abovementioned MSS S and T or copies of them, whereas four manuscripts

seem to be in the private possession of the Chahiāt family, Basile Chammā, Dimitri Qass Nasrallah and ‘Abd al-Masīḥ al-Baramoussī al-Massoudī Salib. With these last four manuscripts we are about to enter the domain of myths and urban legends. The MSS from the *al-Fihris* may still be in the possession of the heirs of the aforementioned persons, or may have been sold to individual merchants, or perhaps, assuming a worst-case scenario, simply do not exist (any more).

3. The ADP

Although editions of the AD have been available for several decades, the discussion in section 2 above has shown their inadequacies and the difficulty of using them in effectively investigating the text and reconstructing its history.

Moreover, a remarkable tendency by editors—Ciasca (1888) and, even more so, Marmardji (1935)—to normalise the text and even to “correct” its presumed language errors makes it rather complex for the reader to reconstruct the original text.¹⁷

A desire to make the text as accessible as possible to the interested reader was one of the key motives behind the ADP.¹⁸ We shared a common feeling that the Arabic translation is per se an interesting text, but not just from the perspective of reconstructing a hypothetical original Syriac text.

3.1. The general framework

The complexity of the transmission history of the text prompted an entirely different approach from that of existing editions. In particular, no effort was made to establish a lemma or textus receptus; instead we took advantage of the possibility offered by digital encoding to make all variants on the same basis available to users.

This choice has a number of advantages from the point of view of both encoding and user experience. First, the marking of all the variants extant in the critical apparatus are placed on an equal footing (see 3.3 below) but do not impose any intrinsic ordering on the encoding of sources and further allows researchers to choose how to proceed according to practical and empirical issues (such as the availability of sources, the amount of editorial work anticipated and so on). As a fallout of this choice, preference can be given to availability over accuracy, e.g., encoding variants in manuscripts according to printed sources and later checking the accuracy of their representation,

since editing single variants does not imply a full rewriting of the apparatus, as in the practice of producing an ordinary critical edition.

Another advantage from the point of view of editors derives from the elimination of the two-stage process, according to which sources are first examined in order to determine the base text before the start of the actual editorial work; since the final encoding will be the same no matter which source is regarded as the “best one”, researchers can start in any order they choose. This choice is also a time-saving element during the encoding phase: if for whatever reason the original selection of the base text proves wrong, no major adjustments are needed.

This is not just a theoretical issue; in the actual encoding process, we understood—contrary to most current learning—that the older, often disliked Ciasca edition (Ciasca 1888) was much better at recording variants and not necessarily worse at the representation of the base text than the often praised newer edition (Marmardji 1935). In a traditional setting, a major reworking of the entire project would have been in order; in the ADP, we just kept encoding.

From the point of view of user experience, the lack of a “preferred reading” makes switching from one reading to another extremely simple: the user just selects a different reading or a set of different readings to compare, and the system performs a quick query that provides relevant information to the user (which normally the editors would have provided).

3.2. *General encoding issues*

The encoding of the AD strictly follows the guidelines of the TEI standard version P5 (The TEI Consortium 2014). Contrary to (some) current practice, we decided to encode texts and variants directly in TEI-compliant XML format, rather than using some intermediate, third-part format. Several reasons were taken into consideration when making this choice:

1. Classical edition software, such as the Classical Text Editor,¹⁹ forces encoders to follow a specific workflow—typically orientated towards the needs of the Classical scholars and not particularly well-suited to tackle the specifics of Medieval Arabic texts—and limit encoding to a subset of the TEI standard. In this context, TEI is only an export format, not the native encoding of the edition;

2. While offering more comfortable GUIs (Graphical User Interfaces), graphical edition software makes the link between text and encoding less transparent and lets subtle errors slip into the encoded text;
3. XML + TEI validation is not too difficult to understand once an encoder is properly taught and encoding standards are documented well enough;
4. The text-only nature of XML means that encoders can work with different tools and editors while keeping maximum compatibility.

As a consequence, we let encoders use their preferred XML editor, with a preference for free editors (for both budget-oriented and philosophical reasons), in order to avoid tagging errors and to ensure compliance with the TEI standard. Consistence of encoded files was enforced through periodical validation against the TEI DTDs.²⁰

Unicode was chosen from the very beginning as the sole reasonable option to encode native Arabic texts natively. While mixing TEI tags and Arabic text can sometimes decrease the readability of the encoding, experience has showed that using transliteration systems based on the Latin alphabet (such as Tim Buckwalter's system) is a very demanding task for encoders and produces much more errors.²¹

3.3. *Encoding of the critical apparatus*

The critical apparatus is being encoded by using only the apparatus (<app>) and reading (<rdg>) tags in-line, together with the reading group tag (<rdgGrp>) to categorise apparatus entries according to subvariation. According to the general editorial principles outlined above (3.1), the lemma (<lemma>) tag was entirely discarded, and as a consequence, all the apparatuses were recorded in-line with the text, rather than encoding it as an external apparatus linked to the text through a location reference.

As to the encoding of witnesses, the ADP adopted a rather innovative approach by drawing inspiration from Marmardji's idea of including the text of the Ciasca edition as an additional witness, marked as 'Ce' in the apparatus of Marmardji edition (Marmardji 1935, p. vi). We marked variants in the text of the two editions by adding "e" to the initials of the editors' surnames—Ce for Ciasca edition and Me for Mamardji edition—while marking manuscripts with a starting sharp sign (e.g., "#A" for the Vatican Arabo manuscript, see 2.1, or "#B" for Vatican Borgianum, 2.2).

The idea to include the editions together with true witnesses, while debatable, is quite plausible for a number of reasons. First, references to the AD text in the academic literature are to editions (mostly to the Marmardji edition), and failing to include them in a new digital edition would make cross-referencing exceedingly cumbersome.

Another good reason to include the two editions is linked to translations. The AD has been translated several times, in Latin (Ciasca 1888), English (Hamlyn Hill 1894; 1919; Hogg 1897; 1906) and German (Preuschen 1926) from the Ciasca edition, and in French (Marmardji 1935) by the editor Marmardji himself. While no translation can be regarded as perfect, they are the sole way to access the text for readers who do not know Arabic (and most scholars interested in the Diatessaron actually do not), and of course, even scholars who know Arabic often find it useful to know how others interpreted a text that is not always easy to read. Since the ADP aims to reach as many scholars as possible, the inclusion of existing translations is absolutely needed, and each translation, in order to be useful, must be linked to the text it actually translates.

The editorial work started by digitalising the Ciasca edition and then collecting variants in the Marmardji edition and in the manuscripts cited in the two critical apparatuses (A and B in both editions, and E only in Marmardji edition). Every time a variant is recorded, the reading of the five witnesses (in the sense just stated) is marked, as far as it can be detected in the two printed editions.²²

Fig. 1: Encoding of orthographic variants²³

In

Fig. 1 we are able to see a sample of how a variant is reported. In this case (a quotation from Joh 1:1), all witnesses agree with the reading **البدء** except for manuscript B, which has an orthographic variant **البدو**. In agreement with the general principle that no reading is marked as a base text, all forms are marked as a reading, with the nonstandard orthography marked with an attribute “type” defined as “ortho” (for “orthographic”). These variants are usually marked in the critical apparatus by Ciasca only; Marmardji disregards them as unsubstantial. However, a good reason to include them is that they may reveal crucial phonetic phenomena that are central to the definition of Middle Arabic, the variety of substandard Medieval Arabic that surfaces often in written documents and that reflects specificities of local spoken varieties. Apart from a specific linguistic interest, this feature might help locate the origin of the copyist (Blau 1965; 2002).

How subtle the distinction between orthographic and substantial variants can be is shown in Fig. 2 below:

Fig. 2: A seemingly orthographic variant

In the sample in Fig. 2, we are presented with a seemingly orthographic variant, discarded by Marmardji and recorded in Ciasca apparatus only as **الله**; *Allāh* or “God” (lit. “the God” with an article) is written with a single “l” **له**. In fact, this “error” (admittedly quite strange in a religious text) can be read as *ilāh* or “God” without an article and reveals itself as one of the most important

substantive variants in the prologue of the Gospel of John quoted in the sample (Kachouh 2007).²⁴ We marked the variant accordingly as “subst” (for “substantive”).²⁵

4. The database and the website

A theoretically preferable strategy to serve TEI-compliant, XML-encoded, critically edited texts to users is to produce (X)HTML files through content-selecting XSL transformation and formatting CSS directives. This workflow is very neat—if, as we actually did and as should happen in any case to match standardisation needs, XHTML is chosen for the final output, everything keeps being encoded in XML throughout the process—and can work very well for small projects, as in the digital edition of the Washington Manuscripts of the Epistles of Paul (Finney 2006).

However, we chose not to follow this workflow for several reasons. First, writing and managing XSL transformations is a notoriously cumbersome process, due to the declarative nature of XSL and the difficulty, notwithstanding the elegance of the project, of “talking about XML in XML”. In fact, as soon as the encoding becomes too complex, writing correct transformations thereof and detecting errors becomes a nightmarish experience. The current available tools are as yet not able to configure the information in order to make the search experience more rewarding.

Second, and perhaps the most important point, is the performance in the current XSL rendering engines, which degrades ungracefully (and sharply) when applying XSL transformations to larger XML source files. The only choice to make a system based on the “natural workflow” working is to divide the source file into a series of fragments, which can undergo transformations with a relatively quick output.

However, this practice is not very sensitive. On the one hand, dividing a meaningful, unified TEI-compliant XML file in a number of fragments makes the whole enterprise somewhat botched. After all, one big advantage of TEI is the possibility of taking a text and annotating it with any sort of meaningful structural information without altering the text itself; if the text is more or less artificially chunked, the utility of encoding it in TEI is called into question.

More specifically, one defining character of the AD (indeed, of the very idea of a Gospel harmony) is its multi-layered structure. The only original segmentation of the text is its subdivision in *iṣḥāḥs* or chapters—verses are lacking from manuscripts, as is usual in New Testament texts from Late Antiquity and the Early Middle Ages, while most modern editions and translations, with

the exception of Ciasca (1888), follow the verse numbering in Hill translation (Hamlyn Hill 1894; 1910) for practical and somewhat questionable reasons. Thus, the possibility of following the layer of quotations of Gospel verses is a perfect alternative. Crucially, quotations do not only cut through verses (in a way editors do not even agree with), but also through chapters. Serving chapter-segmented chunks of the AD text would entirely exclude the possibility of this parallel reading.

4.1. *The database*

We preferred an alternative, more performing and only slightly less elegant solution instead: automatically generating a MongoDB database that records all information in the XML source file. MongoDB is a non-relational database that replaces the usual concept of the database record with the more flexible, especially text-oriented, concept of the document.²⁶ MongoDB documents are encoded internally in BSON, a binary version of the JSON encoding. JSON is the most important alternative to XML for encoding data in a human-readable (actually much more readable than XML), exchange-oriented form.

While some philosophical details in XML and JSON do differ, the two standards are close enough to allow automatic conversion tools to convert from one another in a perfectly acceptable way (Boyer, et al. 2011), especially when—as happens in this case—only a one-way transformation (from XML to JSON) is needed, and the databases is used only in queries (i.e., it is read only and never edited).

Here is how the revised workflow works: every time a new version of the XML source is released, it is automatically transformed into the JSON format, and the MongoDB database is then updated. User queries continue to read the MongoDB database seamlessly, without any change to the user interface.

4.2. *The website*

The public interface to the ADP is hosted at the domain diatessaron.org. When the project expands to versions of the text in other languages, the contents will be moved to the subdomain arabic.diatessaron.org (which now exactly mirrors the contents of the diatessaron.org). A screenshot of the website's welcome page is reproduced in

Fig. 3.

Fig. 3: Screenshot of the welcome page of diatessaron.org

Access to the website is open and free of charge to researchers, upon filling out a registration form. We are currently ensuring that the website has facilities to allow users to provide their feedback for the project managers so that such users' comments can be used to improve access and correct any typos and other errors.

We are currently evaluating the pros and cons of adding a User Wiki as a forum to discuss any issues regarding the electronic edition and to provide forms of cooperative editing. As is well known, cooperative editing in scholarly projects raises a number of issues (quality of contributions, clashes among different theoretical and philosophical points of view and so on), which must be accurately weighted before starting a true Wiki-style collaborative platform.

The website—managed primarily by Giuliano Lancioni—is designed to be mobile-first, so that it can be visualised reasonably well on tablets and even smartphones via automatic resizing and

rearranging of visual elements. This design choice, based upon the adoption of the Bootstrap library,²⁷ addresses the growing diversity in access strategies to information.

The current stage of the project includes the texts of the two published editions, Gospel sources according to Ciasca edition, paging of both editions, together with variants in the A, B and E manuscripts as recorded in the two critical apparatuses. The following steps will include

1. adding Hamlyn Hill's numbers for verses according to the segmentation in the Marmardji edition (which marks the beginning of verses in both the Arabic text and the French translation);
2. linking existing translations to the editions;
3. checking variants from the actual manuscripts and adding more manuscripts not used in previous editions;
4. lemmatising the Arabic text and its variants.

Lemmatisation is a crucial aspect of the ADP; without assigning words to lemmas, search would be severely limited, since only substrings of actual word forms would be returned. Since Arabic has a complex morphology characterised by discontinuous constituents, such a limitation would amount to missing many reasonable results.

In this regard, lemmatising Arabic texts involves two distinct aspects, on both the theoretical and the empirical side. The main theoretical issue is linked to the difficulty to define lemmas, since there has been a significant degree of disagreement between different dictionaries, both Arabic and Western, in what constitutes a lemma (an issue intertwined with the status of the root in Arabic).

On the empirical side, there is actually no truly working lemmatiser for Classical Arabic texts. The most used Arabic lemmatiser is based upon Tim Buckwalter's morphological analyser, AraMorph, which has many shortcomings even for the Modern Standard Arabic texts, is tailored for and is plainly inadequate for Classical Arabic (Buckwalter 2002).²⁸

In order to overcome these difficulties, our research team is working on two different perspectives: the definition of lexical entries and the design of a new lemmatiser.

As to lexical entries, our aim is to detect the optimal solution to enhance search results, rather than to meet some theoretical lexicographical standard. Such an optimal solution is supposed to maximise relevant results while taking into account the specificities of the AD text.

Since the Diatessaron, as a Gospel harmony, is chiefly a compilation of excerpts from the Canonical Gospels,²⁹ a lemmatisation that is linked to standard NT (New Testament) lexicographic

tools would be most useful to the users. For this reason, we started developing a base glossary that links lemmas (according to Classical Arabic dictionaries) to Strong numbers assigned to NT lexical items. Parallel to this point, we are keeping lexical entries as small as possible by disambiguating POS (part-of-speech) categories that tend to conflate in standard Arabic dictionaries, such as nouns and adjectives.

On the implementation level, we are developing a new morphological analyser and lemmatiser based upon Salmoné's Arabic-English dictionary (Salmoné 1890). The reason for this choice is linked to the use of English as a target language (while most Classical Arabic bilingual dictionaries historically preferred Latin or French) and its free availability in a digital edition in TEI-compliant XML format.³⁰

5. Appendix: Sigla, Description and Dating of the Manuscripts of the AD and the Related Coptic-Arabic Gospel Harmonies

5.1. *Family A (or: Antioch Group)*

Z: Lost archetype: Codex antiquus Antiochenus (by Abū 'l-Faraj Ibn al-Ṭayyib based on 'Īsā Ibn 'Alī al-Mutaṭabbib's Syriac text; probably tenth century A.D.).

A: Vatican Arabo XIV (probably thirteenth century A.D.).

C1: Bibliothèque Orientale, Beirut 429/Beirut Fragments [three stray folios only] (fourteenth century: 1332 A.D.) by Abū 'l-Barakāt Ibn abī 'l-[Kabar?]. Could the latter perhaps be a younger relative of Abū 'l-Barakāt Ibn Kabar (d. 1324 A.D.)?

C2: Lost Diatessaron by Yūḥannā Ibn al-Mu'taman *al-ma'rūf bi-Ibn al-Ṣayḥ* (probably thirteenth century A.D.).

C3: Lost Diatessaron by Sim'ān [Ibn Kalīl Ibn Maqāra Ibn Abī 'l-Faraj?] (probably early or mid-thirteenth century A.D.).

C4: Lost Diatessaron by Bishop Yūsāb of Fuwwah a.k.a. Ibn al-Muḥabrik (around 1235-6 A.D.).

5.2. *Family B (or: Baḡdād Group)*

Y: Lost archetype by [?] (probably eleventh century A.D.).

B: Vatican Borgianum Arabo 250 (sixteenth century; post circa 1517 A.D.).

E: Coptic Orthodox Patriarchate, Cairo, *muqaddasa* 136; *raqm musalsal* 131 (eighteenth century: 1795 A.D.). This MS was previously known under the incorrect number COP Cairo no. 202.

O: Bodleian Library, Oxford. Ms. Arab e 163 (early nineteenth century: 1806 A.D. Its lost predecessor [MS X] was supposedly composed by members of the Awlād al-‘Assāl in the early twelfth century: 1107 A.D.).

S: Paul Sbath Library, Aleppo 1020 (late eighteenth century: 1797 A.D.) = Salem Ar. 218.

T: Paul Sbath Library, Aleppo 1280 (probably late eighteenth century A.D. or even early nineteenth century A.D.) = Salem Ar. 446.

Q: Coptic Orthodox Patriarchate, Cairo, *muqaddasa* 135; *raqm musalsal* 198 (eighteenth century A.D.).

5.3. *Other Arabic Gospel Harmonies*

CH1: Paul Sbath Library, Aleppo 1029: *Kanz al-abrār al-aḥyār* by an anonymous Coptic author (perhaps Sim‘ān [Ibn Kalīl Ibn Maqāra Ibn Abī ‘l-Faraj?]) in the early or mid-thirteenth century A.D. = Salem Ar. 227 (del Río Sánchez 2008, p. 127).

CH2: Paul Sbath Library, Aleppo 1038: *Qawānīn al-injīl* by the Coptic author/copyist Abū Sa‘īd ibn Sayyid (or: Sīd) al-Dār Ibn Abī -‘l-Faḍl al-Masīḥi (early fourteenth century: 1322 A.D.) = Salem Ar. 236 (del Río Sánchez 2008, p. 132).

Notes

¹ This paper is the result of joint work. However, the authorship can be attributed as follows: 2 and 5 have been written by Joosse; 1, 3 and 4 by Lancioni.

² There is de facto no communis opinio on the issue of the original language of the Diatessaron. Petersen (1994) opted for a Syriac original, whereas Baarda (2012) pleaded for a Greek original, which in our view is correct. Later on, Syriac translations were produced on the basis of the Greek original. Unfortunately, the Greek and/or Syriac texts of AD have not been preserved for posterity.

³ St. Ephrem’s commentary has long been known in a later Armenian translation. The original Syriac text was found in the late 1950s and edited by Leloir, who published also a most useful French translation,

which compares the Syriac text with the Armenian translation where significant differences arise (Leloir 1963; 1966; 1990).

⁴ Cf. Joosse 2012, p. 69-70.

⁵ Cf. Joosse 2012, p. 70-75.

⁶ Electronic and hard copies of all of these texts are kept at the Theological Faculty (Faculteit der Godgeleerdheid) of the Free University of Amsterdam (VUA), The Netherlands.

⁷ Cf. Joosse 1999, p. 80-81.

⁸ Cf. for the editions of the Arabic Diatessaron (Joosse 1999, p. 86-91). For the different translations of the Arabic Diatessaron, cf. Joosse 1999, p. 91-8.

⁹ Cf. Joosse 1999, p. 81-2.

¹⁰ Cf. Joosse 1999, p. 82.

¹¹ Cf. Joosse 1999, p. 82-3; 2012, p. 68-9.

¹² Cf. Joosse 1999, p. 82-3; del Río Sánchez 2008, p. 121.

¹³ Cf. Joosse 1999, p. 83; del Río Sánchez 2008, p. 248-9.

¹⁴ Cf. Joosse 1999, p. 84-5; 2012, p. 76-8.

¹⁵ Cf. Joosse 2012, p. 69-70.

¹⁶ Cf. Simaika and Yassā 1939-42, p. 2/1/83; Graf 1944-1953, p. I, 668, lines 4-7.

¹⁷ From this point of view, Ciasca (1888) is overall more reliable, since his edition reports consistent variants in manuscripts. Marmardji (1935) usually avoids mentioning variants he judges merely formal or orthographic, while spending large amounts of space in correcting what he deems “faults” in the style of the text, as a corroboration of his well-known theory about the wrong attribution of the Arabic Diatessaron to Ibn al-Ṭayyib.

¹⁸ Encoding work is being put forth by a team, directed by Giuliano Lancioni, that includes Raoul Villano, Marta Campanelli, Ilaria Cicola, Simona Olivieri, Ivana Pepe and Valeria Pettinari.

¹⁹ <http://cte.oeaw.ac.at/>.

²⁰ The choice preferred by most encoders was actually the freeware Notepad++ (<http://notepad-plus-plus.org>).

²¹ An additional benefit deriving from the adoption of Unicode encoding is the possibility of encoding ambiguous, dotless letters in manuscripts. In this respect, Unicode characters ﺀ (U+066E, ARABIC LETTER DOTLESS BEH) and ﺀ (U+066F, ARABIC LETTER DOTLESS QAF), are particularly useful—other dotless letters do exist in ordinary Arabic script, although ambiguity between a properly dotless letter, such as ﺀ, and a dotless version of a dotted letter, such as ﺀ for ﺀ, may arise.

²² While encoding of editions and manuscripts runs in a given order for practical reasons, no version is recorded as a base text. Portions of text are left unmarked only if they are common to all variants,

otherwise they are encoded as belonging to a particular variant. In other words, there is no “textus receptus” in this project, which purposely does not commit to a specific reading.

²³ We thank Ilaria Cicola for helping compose

Fig. 1 and Fig. 2.

²⁴ Giuliano Lancioni thanks Sara Schulthess for pointing him to Kachouh’s article.

²⁵ While in most cases, variant readings are labeled with either the “substantive” or the “orthographical” type, mixed types are possible, in cases where an orthographical variant gives rise to a change in meaning. According to XML attribute syntax, in this case the “type” attribute would have the space-separated value “subst ortho”. Of course, the substantive vs orthographical opposition, while quite clear in the prototypical case, is somewhat blurred sometimes and a decision must be taken by weighing competing reasons for different encodings. In the case of the الله/اله variant, we decided to consider it as an orthographical one after all, since the possible substantive consequences of the variant reading, on a theological or other plane, are not very clear in literature.

²⁶ The MongoDB 3.0 Manual, <http://docs.mongodb.org/manual/>.

²⁷ <http://getbootstrap.com/>.

²⁸ The same applies to more powerful analytical tools, such as MADAMIRA (Pasha, et al. 2014), which are tailored to MSA newspapers and informal texts and perform poorly on Classical Arabic texts, especially when linguistically marginal features, such as those found in the Arabic Diatessaron, are involved.

²⁹ The Diatessaron is not only a compilation of excerpts from the Canonical Gospels; there are also non-Canonical or extra-Canonical elements in AD, as well as variant readings unique to AD. On the issue of non-canonical material in AD, see Petersen (2012).

³⁰ <http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3Atext%3A2002.02.0005>.

Bibliography

Baarda, T., 2012. Tatian’s Diatessaron and the Greek Text of the Gospels. In: Hill, C. E. & M. J. Kruger, eds. 2012. *The Early Text of the New Testament*. Oxford: Oxford University Press. pp. 336-49.

- Blau, J., 1965. *The emergence and linguistic background of Judaeo-Arabic: a study of the origins of Middle Arabic*. Oxford: Oxford University Press.
- , 2002. *A handbook of Early Middle Arabic*. Jerusalem: The Max Schloessinger Memorial Foundation, The Hebrew University of Jerusalem.
- Boyer, J. et al., 2011. *Experiences with JSON and XML Transformations*. IBM Submission to w3C Workshop on Data and Services Integration, October 20-21, Bedford MA. [online] Available at: http://www.w3.org/2011/10/integration-workshop/p/XML_JSON_mapping_paper.pdf [Accessed 01 May 2016]
- Buckwalter, T., 2002. *Buckwalter Arabic Morphological Analyzer Version 1.0*. Philadelphia: Linguistic Data Consortium.
- Ciasca, A., 1888. *Evangeliorum harmoniae arabice*. Romae: Propaganda Fide.
- del Río Sánchez, F., 2008. *Catalogue des manuscrits de la fondation Georges et Mathilde Salem (Alep, Syrie)*. Wiesbaden: Reichert.
- Finney, T. J., 2006. Manuscript Markup. In: Hurtado, L. W., ed. 2006 *The Freer Biblical Manuscripts: Fresh Studies of an American Treasure Trove*. Atlanta: Society of Biblical Literature. pp. 263-87.
- Graf, G., 1944-1953. *Geschichte der christlichen arabischen Literatur*. Città del Vaticano: Biblioteca Apostolica Vaticana.
- Hamlyn Hill, J., 1st ed. 1894; 2nd ed. 1910. *The Earliest Life of Christ Ever Compiled from the Four Gospels being the Diatessaron of Tatian [circ. A.D. 180] Literally Translated from the Arabic Version and containing the Four Gospels woven into One Story*. Edinburgh: T. & T. Clark.

- Hogg, H., 1st ed. 1897 (Edinburgh); 2nd ed. 1906. Tatian's Diatessaron. In: Menzies, A., ed. 1906 *The Ante-Nicene Fathers*. New York: Scribner, Vol. IX. pp. 63-129.
- Joosse, N. P., 1999. An Introduction to the Arabic Diatessaron. *Oriens Christianus*, 83, pp. 72-129.
- Joosse, N. P., 2012. From Antioch to Baġdād, From Baġdād to Cairo: Towards an Archetype of the Arabic Diatessaron. *Parole de l'Orient*, 37, pp. 67-84.
- Kachouh, H., 2007. The Arabic Versions of the Gospels: A Case Study of John 1.1 and 1.18. In: Thomas, D., ed. 2007. *The Bible in Arab Christianity*. Leiden/Boston: Brill. pp. 9-36.
- Leloir, L., 1963. *Saint Ephrem: commentaire de l'évangile concordant: texte syriaque (manuscrit Chester Beatty 709)*. Dublin: Hodges Figgis & Co..
- Leloir, L., 1966. *Éphrem de Nisibe, Commentaire de l'Évangile concordant ou Diatessaron*. Paris: Éditions du cerf.
- Leloir, L., 1990. *Saint Éphrem, Commentaire de l'Évangile Concordant. Texte syriaque (Manuscript Chester Beatty 709). Expanded and Revised Edition*. Leuven: Peeters.
- Marmardji, A.-S., ed., 1935. *Diatessaron de Tatien : texte arabe établi, traduit en français, collationné avec les anciennes versions syriaques, suivi d'un évangeliare diatessarique syriaque et accompagné de quatre planches hors texte*. Beyrouth: Imprimerie Catholique.
- Pasha, A. et al., 2014. MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In: Calzolari, N., et al. eds. 2014. *LREC 2014 Proceedings*. Paris: European Language Resources Association. pp. 1094-1101.
- Petersen, W. L., 1994. *Tatian's Diatessaron: Its Creation, Dissemination, Significance and History in Scholarship*. Leiden, New York, Koln: Brill.

- Petersen, W. L., 2012. Canonicity, Ecclesiastical Authority, and Tatian's Diatessaron. In: Krans, J. & J. Verheyden, eds. 2012. *Patristic and Text-Critical Studies: The Collected Essays of William L. Petersen (New Testament Tools, Studies and Documents 40)*. Leiden, Boston: Brill pp. 489-516.
- Preuschen, E., 1926. *Tatians Diatessaron, aus dem arabischen übersetzt von Erwin Preuschen ; mit einer einleitenden Abhandlung und Textkritischen Anmerkungen herausgegeben von August Pott; unter Beihilfe der Heidelberger Akademie der Wissenschaften*. Heidelberg: C. Winters.
- Salmoné, H. A., 1890. *An Arabic-English Dictionary on a New System*. London: Trübner.
- Sbath, P., 1938. *Al-Fihris : catalogue de manuscrits arabes*. Le Caire: s.n.
- Simaika, M. & Yassā, ' . a.-M., 1939-42. *Catalogue of the Coptic and Arabic Manuscripts in the Coptic Museum, the Patriarchate, the Principal Churches of Cairo and Alexandria and the Monasteries of Egypt*. Cairo: Govt. Press.
- The TEI Consortium, 2014. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. s.l.:Text Encoding Initiative Consortium.