First Steps Towards the Integration of Resources on Historical Glossing Traditions in the History of Chinese: A Collection of Standardized Fǎnqiè Spellings from the Guǎngyùn

Michele Pulini[1,2], **Johann-Mattis List**[2,3]
Ca' Foscari University[1], University of Passau[2], Max Planck Institute for Evolutionary Anthropology[3],
Venice, Italy[1], Passau, Germany[2], Leipzig, Germany[3]
michele.pulini@unive.it, mattis.list@uni-passau.de

Abstract

Due to the peculiar nature of the Chinese writing system, it is difficult to assess the pronunciation of historical varieties of Chinese. In order to reconstruct ancient pronunciations, historical glossing practices play a crucial role. However, although studied thoroughly by numerous scholars, most research has been carried out in a qualitative manner, and no attempt at providing integrated resources of historical glossing practices has been made so far. Here, we present a first step towards the integration of resources on historical glossing traditions in the history of Chinese. Our starting point are so-called *fǎnqiè* spellings in the Guǎngyùn, one of the early rhyme books in the history of Chinese, providing pronunciations for more than 20000 Chinese characters. By standardizing digital versions of the resource using tools from computational historical linguistics, we show that we can predict historical spellings with high precision and at the same time shed light on the precision of ancient glossing practices. Although a considerably small first step, our resource could be the starting point for an integrated, standardized collection that could ultimately shed new light on the history of Chinese.

Keywords: fǎnqiè spellings, Chinese Historical Phonology, history of Chinese

1. Introduction

Historical glossing practices by which the pronunciation of a character was elucidated are a particularly interesting aspect of Chinese Historical Phonology. As it is well known, the Chinese writing system gives only minimal hints regarding the pronunciation of its characters. A character like 手 'hand', pronounced as *shǒu* (or [ʂɔu^{214}] in the International Phonetic Alphabet, hereafter IPA), does not tell us anything about its pronunciation today and even less about its pronunciation in the past. Since the ancient Chinese scholars didn't have an alphabet to simply transcribe their sounds (Sino-xenic alphabetic transcriptions and intensive contact with Indian phoneticians started much later), they started from simple equations according to which one character was associated with a homophonous or phonetically similar character. The two methods are often referred to as *zhíyīnfǎ* 直音法 (direct reading method) and the *bǐnǐfǎ* 比擬法 or *dúruòfǎ* 讀若法 (similar reading method) (Coblin, 1983). This latter denomination comes from the formula used in some lexicographic works to suggest the reading of an entry: e.g. in his *Shuōwén Jiězì* 說文解字 (written between 58 and 157 CE), Xǔ Shèn occasionally uses the formula "read [this character] as X" (*dúruò X* 讀若《丙》) in addition to his explanations of the meanings and the structure of the characters. The disadvantage of the *zhíyīn* and the *dúruò* methods is that they only allow to gloss characters for which a common character with an identical or a very similar pronunciation exists. It is also not clear to what extent certain deviations were allowed.

The so-called *fǎnqiè* spellings (Coblin, 1983; Branner, 2000) provided a much more precise way of glossing character pronunciations. This spelling method, which seems to go back to at least the third century, was born as a practical response to the necessity of transliterating new vocabulary of Indic origin due to the spread of Buddhism in the Central Plains (Sun and Wu, 2015). The system is based on breaking the character pronunciation into two parts, the *initial* and the *final*, and selecting two different glossing characters, one with an identical initial sound and one with the identical final. Figure 1 (modified from List 2018) illustrates this technique for the character *dōng* 東 "east".

Given their straightforwardness and simplicity, *fǎnqiè* pronunciation glosses became quite popular among Chinese scholars, and even today, people may occasionally use them in order to explain pronunciations without having to rely on foreign writing systems like the Latin alphabet. As a result, there is an abundance of sources which use this pronunciation device throughout the history of the Chinese language. Although the pronunciation is only given indirectly, with respect to the pronunciation traditions which were active at a given epoch, the *fǎnqiè* spellings offer great help to explore how the pronunciation of the Chinese language changed over time.

LREC-COLING 2024, pages 7343–7348

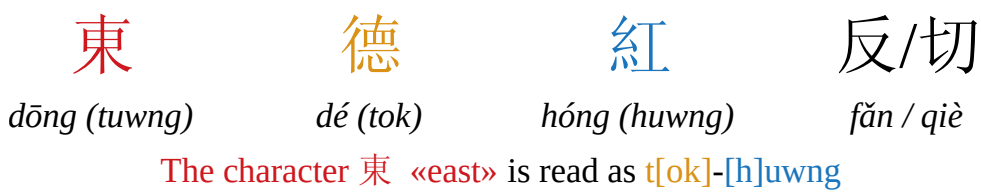

Figure 1: The *fǎnqiè* spelling method.

Most of the research on pronunciation changes investigated through the comparison of *fǎnqiè* spellings has been carried out manually. Chinese scholars' systematic work on *fǎnqiè* spellings goes back to the early 19th century, when scholars began to investigate which characters were used to denote certain initial sounds (*fǎnqiè shàngzì* 反切上字), and which characters were used to denote the finals (*fǎnqiè xiàzì* 反切下字). Although a maximum of two spellers per single initial and final could have been sufficient for handling the totality of the spelling task, the Chinese lexicographers often employed many different characters for the same phonetic values. Fortunately, the alternations were more or less consistent, with some characters being used more frequently and some characters being used less frequently. For example, we can say that the characters *gōng* 公, *gǔ* 古, *gàn* 干, etc. were regularly used to indicate initials which would be spelled as [k] in the IPA, while *kǒu* 口, *kě* 可, and *kǔ* 苦 were used to pronounce [k^h].

Detailed manual and in part even quantitative investigations of *fǎnqiè* allowed scholars to reconstruct different stages of Chinese that are now broadly known as *Middle Chinese*, an abstract variety dating back to the language encoded in various collections of *fǎnqiè* spellings published from about the 6th until the 12th century.

While individual investigations of these spelling collections have greatly helped to elucidate the historical development of Chinese varities, no attempts have been made to systematically integrate and standardize these resources. Various databases that include *fǎnqiè* spellings and other historical techniques for the glossing of pronunciations have been published, but – to the best of our knowledge – no attempt has been made to provide a way to unify digital resources and to standardize them. Here, we propose a first step towards the integration of resources on historical spelling traditions by providing a new collection of standardized *fǎnqiè* spellings taken from the Guǎngyùn 廣韻 (*Broad Rhyme Collections*), one of the largest historical collections of character pronunciations published in 1007/1008 CE. Our collection is accompanied by computational tests that allow us to check how well the spellings were annotated by us. In addition, we offer straightforward interactive web-application that allows scholars to predict Middle Chinese character pronunciations from observed *fǎnqiè* spellings in the popular Middle Chinese reconstruction system by Baxter (1992).

2. Materials

We start from the version of the Guǎngyùn edited by Zhōu (1938). The Guǎngyùn itself builds on and extends the Qièyùn 切韻, compiled in 601 CE, which was the first rhyme dictionary in the history of Chinese known to us. The Guǎngyùn in the edition of Zhōu was digitized in a research project funded by the Japan Society for the Promotion of Science and publicly shared in the form of an XML file (https://kanji-database.sourceforge.net/dict/sbgy/). Although data are shared under permissive licenses for this digital resource, the editors of the data are not named in the resource itself.

While Zhōu commented on the Guǎngyùn in his edition, no direct pronunciations in form of alphabetical spellings were proposed for the *fǎnqiè* glosses. In the digital version, however, explicit pronunciations in a rough version of the IPA were added by the editors, supposedly following the schematic nature of the Guǎngyùn and the comments by Zhōu. In our treatment of the data, we attribute these pronunciations to Zhōu, who gave his own indications on the sounds of Middle Chinese based on the Guǎngyùn in other contexts (see Zhōu 2004), but we emphasize that the readings were not provided in the original edition of the rhyme dictionary. We parsed the XML in order to extract all *fǎnqiè* glosses in tabular form.

In addition to the Guǎngyùn in digital form, in order to test the prediction of character readings based on the *fǎnqiè* glosses, we made use of the Middle Chinese readings listed along with Old Chinese reconstructions in Baxter and Sagart (2014). These data, which the authors originally shared in the form of an Excel file, are now available as an appendix to the Wiktionary resource (https://en.wiktionary.org/wiki/Appendix:Baxter-Sagart_Old_Chinese_reconstruction, retrieved on October 17, 2023, last edited on July 9, 2021, at 10:24).

3. Methods

3.1. Annotation of Glossing Characters

Rhyme dictionaries are the largest source of *fǎnqiè* spellings in the history of Chinese. In contrast to other sources, in which *fǎnqiè* spellings are used in an ad-hoc fashion, they provide a systematic collection of *fǎnqiè* characters which can be easily extracted in tabular form and annotated. Based on the digital source of the Guǎngyùn rhyme book, we extracted all *fǎnqiè* spellings for initials and finals in tabular form along with the pronunciations in IPA (that were added by the editors of the digital version, based on annotations by Zhōu 1938).

The syllable of Chinese varieties can typically be divided into five parts, the *initial*, the *medial* (a glide), the *nucleus* (vowel), the *coda* (consonant), and the *tone* (Wang, 1996). With *fǎnqiè* spellings divided into two parts, one providing hints on the initial part of the syllable, and one providing hints on the final, the target pronunciation is usually assumed to be derived from the *initial* and the *medial* of the first character and the *medial*, *nucleus*, *coda*, and *tone* of the second character.

While there is generally consistency in how both upper spellers and lower spellers annotate the medial glide (Jacques, 2015) , in some cases, only the lower speller conveys it: e.g. *luàn* 亂 MC (Middle Chinese) *lwanH* is spelled as 朗 MC *langX* and 段 MC *dwanH*, where only the second character gives a hint on the glide *w*. To account for this, we annotated all upper and lower *fǎnqiè* characters in the data manually, providing initial and medial for the upper and medial, nucleus, coda, and tone for the lower spelling character. Taking inspiration from the technique of template alignments for historical language comparison presented in Wu et al. (2020), we provide pronunciations in IPA for the original readings provided with the digital version of the Guǎngyùn and align them with the Middle Chinese reconstruction system by Baxter (1992), using the EDICTOR, an interactive, web-based tool for the creation and curation of etymological data (List, 2017, 2023). EDICTOR is merely used to code the data and handle phonetic alignments. Its alignment functionalities, in fact, allow us to check for correspondences across the two transcription systems (i.e. Baxter 1992 and Zhōu 1938). Providing readings for character combinations, however, is something EDICTOR was not designed to do.

The annotation is illustrated in Figure 2, where two character glosses are shown, one upper gloss (denoting initial and medial glide) and one lower gloss (denoting the whole rhyme of a syllable). For the upper gloss, we provide initial and medial for both reconstructions by Baxter and by Zhōu. For the lower gloss, the coda position (position three in the template in the last column of the table) is left empty, indicated by a pink square in the tool. As can be seen from the figure, tones and other sounds are represented differently in both reconstruction systems. By aligning the data to the virtual templates representing the basic structure of the Chinese syllable, we can easily compare differences across reconstruction systems.

3.2. Middle Chinese Reading Prediction

Annotated in aligned form, providing two distinct reconstructions for Middle Chinese, we can use the *fǎnqiè* glosses directly to predict Middle Chinese pronunciations for individual Chinese characters. While the prediction of character readings from *fǎnqiè* is straight-forward in theory, in practice, we have to deal with three major issues, (1) multiple pronunciations for the same character in Chinese, (2) potential conflicts in the handling of medial glides, and (3) individual peculiarities of reconstruction systems for Middle Chinese by different scholars.

As an example for (1), consider the character 角 (used as rhyme speller in the Guǎngyùn), it is given two Middle Chinese readings, namely *kaewk* and *luwk*, and can still be pronounced as *jiǎo*, *jué*, *lù*, and *gú* in modern Mandarin. As a result, *fǎnqiè* glosses can have inherent ambiguities, resulting from characters with multiple pronunciations. Regarding (2), the upper glossing character that denotes primarily the initial in *fǎnqiè* glosses can occasionally also give hints on the pronunciation of the medial glide. However, due to the fact that *fǎnqiè* were not designed as a systematic spelling system, but rather evolved as a practice over time, we can encounter conflicts in which the upper gloss shows a medial that is different from the medial in the lower gloss. As to problem (3), scholars who propose Middle Chinese reconstructions from the study of rhyme books and additional sources of historical phonology of Chinese often add peculiarities to their reconstruction systems in order to yield a more economic representation. Thus, while scholars agree that many syllables have a palatal glide as a medial, Baxter (1992) discards the notation of a medial glide after palatal initials, which are already spelled with a *y* in his transcription of Middle Chinese (i.e., *tsy-*, *tsyh-*, *dzy-*, *ny-*, *sy-*, *zy-*, *y-*). Palatal initials only occur with finals beginning with *-j-* or *-i-*, so discarding the notation of a medial glide *j* does not entail a loss of contrast. On the other hand, the medial *w* is not notated after labial initials - except some distinct cases - (*p-*, *ph-*, *b-*, *m-*) since it is not contrastive. However, the presence of a medial *j* or *w* is in most cases conveyed by both the initial and rhyme spellers in *fǎnqiè* glosses.

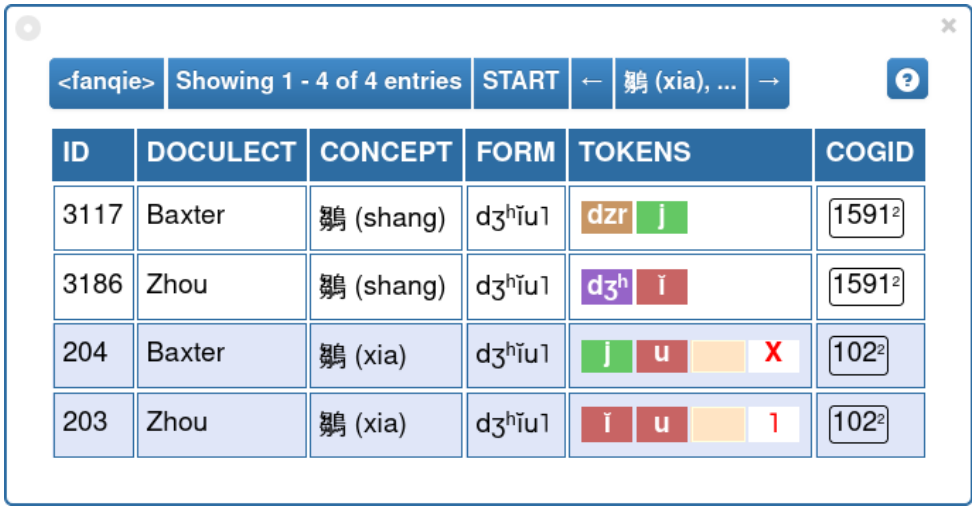

Figure 2: Annotating upper and lower *fănqiè* spelling characters with the help of the EDICTOR tool. The examples shows the character 鶵 *chú* "chick" (variant of 雛), which occurs both as initial and as rhyme speller in the Guăngyùn. The first two rows show the treatment of initial spellers, for which we give the initial and the medial of the characters Middle Chinese reconstruction in the system by Baxter and the system by Zhōu. The two rows below show the rhyme speller annotation, where we provide readings for medial, nucleus, coda, and tone in the two systems. The column names are the general names available in the tool. Column *DOCULECT* distinguishes the two reconstruction sources, *CONCEPT* displays the character and speller type (initial or rhyme), and *TOKENS* provides a segmented reconstruction, aligned against the template (2 and 4 segments).

In order to address these problems, we provide a simple, explicit workflow, by which character pronunciations can be predicted for *fănqiè* glosses following different reconstruction systems. The starting point of our prediction is the combination of the two readings. In a second step, we check if the value for the medial position is identical or conflicting. In the case of a conflict, we use the medial in the rhyme speller. In a third step, we apply a set of simple character replacements to account for peculiarities in reconstruction systems.

3.3. Evaluation

We test our data with the help of a couple of automated tests. These tests help us to avoid errors resulting from manual annotation and to assess to which degree the *fănqiè* gloss collections of the Guăngyùn reflect Middle Chinese reconstructions provided independently of the rhyme book as a major resource.

3.4. Implementation

With the data primarily being curated in the EDICTOR tool, which is usually used for historical language comparison, we offer two implementations, by which *fănqiè* glosses in our resource can be actively used and investigated. First, we offer a little collection of Python scripts that we use to curate and check the data and to carry out predictions from within Python scripts. Second, we offer a light-weight web-based application in JavaScript that allows to insert *fănqiè* glosses into a text field and then yields readings in the Middle Chinese reconstruction system by Baxter and by Zhōu (see Figure 3).

4. Results

4.1. General Statistics

Our resource provides pronunciations for two reconstruction systems for 472 upper glossing characters and 1184 lower glossing characters. Using our prediction method, we can predict Middle Chinese readings for all 24922 characters listed in the digital edition of the Guăngyùn in two reconstruction systems for Middle Chinese. Due to the generative character of *fănqiè* glosses, more readings can be generated, and the resource can be actively tested to investigate *fănqiè* collections from different sources.

4.2. Internal Consistency

We test the internal consistency of our resource by iterating over all 3729 distinct *fănqiè* combinations in the digital Guăngyùn edition and testing to which degree our prediction method predicts the pronunciation provided by the editors. We find that our automatically derived pronunciations are identical in 96% of all cases, with 167 cases where the predicted pronunciation differs from the one proposed in the digital Guăngyùn. Most of these very few cases are due to inherent ambiguities in

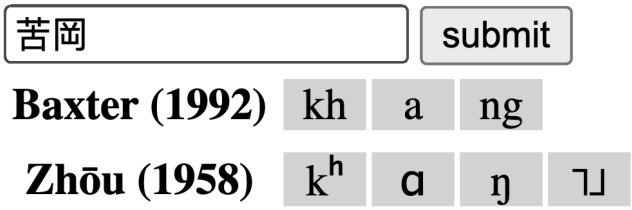

Figure 3: The interactive *fǎnqiè* prediction app.

the underlying readings of *fǎnqiè* glosses. Thus, given as input the characters 丁刮, our prediction proposes *twat* as the reading, while the digital edition supposes *ʈwat* (with [ʈ] representing a retroflex sound). When checking for different readings of the character 丁 in the Guǎngyùn, we find indeed that it has two spellings, namely 當經, transcribed as *tieŋ⟨I⟩* in Zhōu's reconstruction and *teng* in Baxter's system, and 中莖, transcribed as *ʈœŋ⟨I⟩* in Zhōu's system and as *treang* in Baxter's system. This shows that though being true that *fǎnqiè* spellers are mainly employed only with their basic reading, the inherent ambiguity of the pronunciations of Chinese characters can indeed leave traces in the accuracy of *fǎnqiè* glosses, a finding that has – to the best of our knowledge – so far not been actively discussed in the literature on Chinese Historical Phonology.

4.3. Predicting Baxter's Middle Chinese

In order to test how well we can predict Baxter's Middle Chinese reconstructions from the annotated collection of *fǎnqiè* characters, we tested our prediction on Middle Chinese reconstructions accompanying the Old Chinese reconstructions by Baxter and Sagart (2014). In order to do so, we iterated over all 4083 character readings provided in the Wiktionary version of the data and looked up all possible *fǎnqiè* glosses for each character in the Guǎngyùn. Whenever these glosses could be found (for 231 characters, no equivalent was found in the Guǎngyùn), we then predicted Middle Chinese readings in Baxter's system and compared them directly to the Middle Chinese readings provided by Baxter and Sagart themselves. We find that this approach yields at least one correct reading in 91% of all cases. In 330 out of 3852 characters, none of the proposed predicted readings matched the one provided by Baxter and Sagart. While some of these cases may reflect errors in our resource and warrant a correction after thorough examination, it is interesting to note that quite a few cases we observe show only differences in the tone. Thus, we find 32 cases among the errors where our prediction and the readings by Baxter and Sagart show a difference in traditional *shàng* and *qù* tones. At least in part, this is due to the ambiguity of characters regarding their tonal reading. Thus, the character *shǎo* 少 "small" has two readings MC *syewX* (*shàng tone*) and MC *syewH* (*qù* tone), with both readings being employed to spell rhymes in the Guǎngyùn. Since the Guǎngyùn indicates tones independently of *fǎnqiè* spellings, this means that the information within the rhyme book is consistently represented. It shows, however, that one should be careful when finding *fǎnqiè* spellers outside of the systematic organization provided by the rhyme books. This demonstrates on the one hand, that our system is performing already quite well, on the other hand, it shows that we still need to check all of these cases thoroughly in an updated version of our resource and find a way to handle attested ambiguities in spelling characters.

5. Conclusion

The accuracy of predictions is already very promising and makes our tool useful for research purposes, although it is clear that errors must be thoroughly investigated in the future. More importantly, however, our research can be seen as a proof of concept that could be extended and applied to additional resources, including other rhyme dictionaries and lexicographic works that were published in the history of Chinese, such as the *Jīngdiǎn shìwén* 經典釋文 (6th-century CE Textual explanations of classics and canons), a major work in Chinese lexicography of primary importance for the phonological reconstructions of Middle and Old Chinese, but also for philological inquiry and research on ancient Classics. *Jīngdiǎn shìwén* is, in fact, a very extensive work that gathers phonetic glosses (*fǎnqiè*) from previous traditions. Since the work employs a considerable variety of *fǎnqiè* spellers, often indicating different pronunciations for the same character, the availability of a tool like the one proposed by us would prove to be extremely relevant and useful for researchers both in the fields of Chinese historical phonology and philology. This kind of extension to sources other than the Guǎngyùn could also prove crucial in understanding to what extent Middle Chinese rhyme books and lexicographic works vary in the implementation of the *fǎnqiè* spellings.

6. Bibliographical References

William H. Baxter. 1992. *A handbook of Old Chinese phonology*. de Gruyter, Berlin.

William H. Baxter and Laurent Sagart. 2014. *Old Chinese. A new reconstruction*. Oxford University Press, Oxford.

David Prager Branner. 2000. The rime-table system of formal Chinese phonology. In Sylvain Auroux, E. F. K. Koerner, Hans-Josef Niederehe, and Kees Versteegh, editors, *History of the language sciences*, volume 1, pages 46–55. de Gruyter, Berlin and New York.

Weldon South Coblin. 1983. *A Handbook of Eastern Han Sound Glosses*. The Chinese University Press, Chicago.

Guillaume Jacques. 2015. Traditional chinese phonology. In Rint Sybesma, editor, *Encyclopedia of Chinese Language and Linguistics*. Brill. First published online: 2015.

Johann-Mattis List. 2017. A web-based interactive tool for creating, inspecting, editing, and publishing etymological datasets. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. System Demonstrations*, pages 9–12, Valencia. Association for Computational Linguistics.

Johann-Mattis List. 2018. More on Network Approaches in Historical Chinese Phonology (音韵学). In *The 2nd Li Fang-Kuei Society Young Scholars Symposium*, pages 157–174, Taipei. Li Fang-Kuei Society for Chinese Linguistics.

Johann-Mattis List. 2023. *EDICTOR. A web-based tool for creating, editing, and publishing etymological datasets [Software Tool, Version 2.1.0]*. MCL Chair at the University of Passau, Passau.

Jingtao Sun and Hede Wu. 2015. Fǎnqiè 反切. In Rint Sybesma, editor, *Encyclopedia of Chinese Language and Linguistics*. Brill. First published online: 2015.

William S.-Y. Wang. 1996. Linguistic diversity and language relationships. In Cheng-teh James Huang, editor, *New horizons in Chinese linguistics*, number 36 in Studies in natural language and linguistic theory, pages 235–267. Kluwer, Dordrecht.

Mei-Shin Wu, Nathanael E. Schweikhard, Timotheus A. Bodt, Nathan W. Hill, and Johann-Mattis List. 2020. Computer-assisted language comparison. state of the art. *Journal of Open Humanities Data*, 6(2):1–14.

Zǔmó 周祖謨 Zhōu. 1938. Guǎngyùn xiàoběn 广韵校本 [revised edition of the guǎngyùn]. Digital edition from https://kanji-database.sourceforge.net/dict/sbgy/.

Zǔmó 周祖謨 Zhōu. 2004. *Zhōu Zǔmó wénzì yīnyùn xùngǔ jiǎng yì* 周祖謨文字音韻訓詁講義. Tianjin guji chubanshe.

7. Supplementary Material

All annotated data as well as the code that is needed to replicate this study are curated on GitHub (https://github.com/digling/fanqie) and archived with Zenodo (https://doi.org/10.5281/zenodo.10828130).

8. Acknowledgments

This research was supported by the Max Planck Society Research Grant *CALC³* (JML, https://digling.org) and the ERC Consolidator Grant *ProduSemy* (JML, Grant No. 101044282, see https://doi.org/10.3030/101044282). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency (nor any other funding agencies involved). Neither the European Union nor the granting authority can be held responsible for them. We thank Barbara Meisterernst for encouragement and helpful discussions on the resource.