Cleaning up and Standardizing a Folktale Corpus for Humanities Research

Muiser, I. E.C.; Theune, M.; Meder, T.

published in
Proceedings of the Second Workshop on Annotation of Corpora for Research in het Humanities (ACRH-2)
2012

document version
Publisher's PDF, also known as Version of record

Cleaning up and Standardizing a Folktale Corpus for Humanities Research

Iwe Everhardus Christiaan Muiser[1], Mariët Theune[2] and Theo Meder[3]

[1]Database group, University of Twente, Enschede, the Netherlands
[2]Human Media Interaction, University of Twente, Enschede, the Netherlands
[3]Meertens Institute, Amsterdam, the Netherlands

E-mail: e.c.muiser@utwente.nl

Abstract

Recordings in the field of folk narrative have been made around the world for many decades. By digitizing and annotating these texts, they are frozen in time and are better suited for searching, sorting and performing research on. This paper describes the first steps of the process of standardization and preparation of digital folktale metadata for scientific use and improving availability of the data for humanities and, more specifically, folktale research. The Dutch Folktale Database has been used as case study but, since these problems are common in all corpora with manually created metadata, the explanation of the process is kept as general as possible.

1 Introduction

Recordings in the field of folk narrative have been made around the world for many decades. Storage, annotation and studies of these corpora by digital means however, have only begun just recently. By annotating and creating digital versions of these tales, they are virtually 'frozen' in time, which opens up a window for researchers to compare historical versions of narratives. Many types of studies can be efficiently done by searching and comparing the tales once they are put in a digital framework [1]. By adding metadata such as keywords, dates, and geographical locations to the original texts it becomes easier to search, categorize and navigate through a corpus. An additional advantage of digitization is that the information can now be shared with researchers and other interested people all over the world. This creates wonderful opportunities to analyze and compare the similarities and differences between folktales in various cultures. In general the more metadata is present, the more extensive comparative research can be performed.

For a few decades now, the Meertens Institute in Amsterdam has been collecting folktales. Since 1994 these tales are being digitized and put into the Dutch

Folktale Database (DFDB), which came online in 2004 (www.verhalenbank.nl). Over the years, the descriptive metadata assigned to these tales have undergone several transformations due to periodically changing demands. Data fields have been added to describe and identify the original texts in increasing detail, for instance with an indication of the motifs in the text, the creation and mutation date, and whether the tale is extreme in nature. Today, the database contains roughly 42.000 tales, all of which are provided with a rich set of manually added metadata. Several tens of thousands of texts are still waiting for annotation or digitization, and many circulating Dutch folktales have not even been recorded yet.

The Dutch folktale collection in the DFDB is currently being used within the FACT project[1] to investigate the automatic annotation of folktales with metadata such as genre, language and keywords. The project aims to support humanities research by developing new methods for automatic metadata extraction, classification and clustering of folktales. This is a challenging task because many of the folktale texts have been taken down within the context of oral performance, and can contain vernacular language (including laughter, pauses, hesitations, incomplete and imperfect sentences etc.), dialect, slang, and mixed languages. This often causes traditional natural language processing (NLP) methods to be insufficient.

The manually annotated metadata of the folktale corpus are a very valuable resource for the automatic annotation of folktale texts. Information about the date and place of narration is likely to be of use for automatic language identification, and in experiments on classification of folk narrative genres we have shown that metadata such as keywords, summary, and date can be used to improve performance [2]. However, for the current metadata annotations to be optimally useful for training and testing automatic classifiers, an organized metadata setup has to be present, and the number of errors and inconsistencies in the data fields has to be minimized. For example, our language identification experiments may have been hampered by inconsistent labeling of mixed language documents [3].

Since the annotation of folktales has been done by hand by about sixty different annotators over a large period of time, a fair portion of errors and deviations from the input standard can be expected. Van den Bosch et al. have observed error rates up to 5% in cultural heritage databases [4].

This paper describes the first steps of the process of standardization and preparation of the DFDB metadata for scientific use, improving availability of the data for humanities and, more specifically, folktale research. The paper focuses on the Dutch Folktale Database, but, since these problems are common in all corpora with manually created metadata, the explanation of the process is kept as general as possible. In Section 2, some light is shed on common errors in cultural heritage databases. Section 3 discusses metadata standards. The actual standardization of metadata values of our folktale corpus is described in Section 4.

[1] Folktales As Classifiable Texts, http://www.elab-oralculture.nl/fact

2 Errors and inconsistencies in cultural heritage databases

Many digitized cultural heritage collections like the DFDB contain information that was created manually. Most information has been stored in free text formatted databases, in many respects functioning as a digital encyclopedia, without taking into account the possibilities of digital analysis laying ahead. The free text input left room for freedom of annotation, comments and explanations. This is beneficial for free text searching but disastrous for ordering data and structured search. It makes browsing and visualization a very challenging task. Standardization of metadata is therefore of great importance for the field.

Manual free text input also allows mistakes and inconsistencies to easily sneak in. Van den Bosch et al. [4] discuss three types of common errors in cultural heritage databases. Items with *typing and spelling errors* are unlikely to turn up in search results. *Wrong column errors* occur when the content of database columns has been misplaced or switched. *Content errors* are usually due to wrong, or alternative, interpretations and classifications of corpus items (in our case, folktales). Fixing these errors is a step that can be made after clearing up more generic inconsistencies. In the metadata of our corpus we found the following inconsistencies:

- Deviations from a set standard. In dates, for example, 'February 1st 2012' could be '1 feb. 2012', '01-02-2012' or '2012-02-01'. Names can also have many formats like 'Jan van der Vaart', 'van der Vaart, Jan' or 'J. vd Vaart'.
- Differences in delimiters. Names can be separated by comma's, ampersands, or other characters.
- Addition of comments in divergent formats. A value can be uncertain, causing some annotators to add '?', '[?]', 'UNKNOWN', or putting the complete value between square brackets.
- Capitalization / punctuation variations. Some tale titles end with a full stop while others do not. This is also the case for other values such as tale type and geographical location. Names and geographical locations that need to start with a capital character are sometimes completely written in lowercase.

The paper focuses mainly on fixing these inconsistencies. Correction of actual errors will be addressed in a later stage of the project.

3 Metadata standards and infrastructures

Most cultural heritage databases start out using similar terms such as date, names, keywords, or geographical locations. Yet, it is common that during the database's life span, metadata terms are added and their values become more complex. To prepare for cooperation with, and to prevent bad communication between, other collections around the world [5], it is preferable to comply with standards where possible.

Standards for data and metadata are available in abundance. All these standards have their own (dis)advantages and levels of complexity. For standardization of the DFDB we choose to adopt Dublin Core[2], because it is the most basic and widely accepted standard. An additional plus is that the web-publishing platform of choice for the DFDB, Omeka[3], has Dublin Core as its primary standard. Dublin Core is a classic set of 15 metadata terms which can be used to describe a large range of media resources (web and physical). Dublin Core terms can be interpreted loosely due to their general nature. Mappings of database fields to Dublin Core terms might not always be intuitive due to differences in naming conventions. Before assigning a name, the type of data must be properly analyzed. For the sake of internationalization, the field names have to be defined in English. In case of the DFDB however, terms are defined in Dutch, and can, when literally translated, have a slightly different meaning. This can potentially cause confusion when used in an international context.

All data of the Meertens Institute, including the DFDB, will eventually be made available through the CLARIN initiative [6], which aims for a sustainable data infrastructure to aid interoperability in the humanities, and more specifically, linguistics. CLARIN uses a structured data format for metadata called Component MetaData Structure (CMDI). Datasets have to be converted to this format before they can be accessed through CLARIN. To avoid conversion problems, it is of course best to have data that do not deviate from a strict standard. Conversion tools are available for Dublin Core and other, more linguistically oriented metadata schemes such as OLAC[4] and IMDI[5]. To ensure interoperability, CLARIN makes use of ISOCAT, a framework for defining data categories that comply with the ISO/IEC 11179 family of standards. Here metadata terms and their definitions and restrictions can be registered, or existing terms can be adopted when deemed suitable. The latter is always encouraged to limit the number of terms in ISOCAT.

4 Standardization

In this section we describe the main steps involved in standardization of a corpus. First and foremost, it is important to collect the wishes of the users of the database in question. Users rely on data that can be found and sorted based on all available terms. The only way to facilitate this is to make sure that all new items conform to a strict and properly documented standard before being submitted to the database.

4.1 Dutch Folk Tale Database metadata

The DFDB encompasses a rich set of metadata fields: a total of 29 terms supplement the original text. Annotation and input of folktales in the DFDB was largely

[2] http://www.dublincore.org
[3] http://www.omeka.org
[4] http://language-archives.org
[5] http://www.mpi.nl/imdi/

done by interns and employees of the Meertens Institute. Most metadata fields have been entered in a free text format. Tables 1 and 2 show the terms of the DFDB, including mappings to their future standards.

Dublin Core term	DFDB term	Explanation
4.1. Title	title	Title of the folktale
4.2. Subject	folktale/ATU type	Folktale type code
4.3. Description	text summary	Summary of the text
4.4. Type	source format	Original source type (e.g., book, oral)
4.5. Source	text source	Description of the source
4.6. Relation	-	Empty for future use
4.7. Coverage	region	Geographical information
4.8. Creator	narrator	The person who told the tale
4.9. Publisher	-	Not used
4.10. Contributor	collector	The person that recorded the tale
4.11. Rights	copyrights	Specifies if a text is copyrighted
4.12. Date	date	The date of narration or discovery
4.13. Format	-	Empty for future use
4.14. Identifier	id number	The internal identifier code
4.15. Language	language	The language or dialect of narration

Table 1: A folk tale object's metadata terms mapped to Dublin Core terms

DFDB term (English)	Explanation
literary	Specifies if the text is literary
subgenre	Genre of the tale (fairy tale, joke, etc.)
motifs	Comma separated Thompson motif codes
keywords	Comma separated list of keywords
named entities	Named entities mentioned in the text
remarks	Additional information about the text
corpus	Corpus code
definition / description	ATU information
kloeke georeference	Kloeke georeference code of region
kloeke georeference in text	Locations mentioned in the text
extreme	Specifies if a text is extreme in nature

Table 2: List of original DFDB terms that need to be registered at ISOCAT

For a selection of fields, scripts were written to analyze, and to convert the original free text values into well formatted values.

The *date* data type is one of the most diversely composed values in this database. It ranges from perfectly composed ISO 8601 international standard (YYYY-MM-DD), to Dutch standard (DD-MM-YYYY), to completely textual values like 'Third quarter seventeenth century' and 'stumbled upon on 12 February 2003'. Some values have question marks, square brackets, commas and points in them. Sporadi-

cally the date has been supplemented with information about the era when the story was being told, or when the story took place.

Statistics about observed variations in date formats are shown in Table 3. Before the dates were checked for inconsistencies, they were lower cased and spell-corrected. Square brackets around the date number, month, year or whole date value were removed, values like 'sep.', 'sept', 'sept.', and 'september' were changed to 'm09', and day names were taken out as well. Some implausible values like 'February 30th 1969' were recognized by the scripts and manually corrected before the final conversion.

The *region* column is another value that can deviate a fair amount from any defined standard although the vast majority has a 'place (province)' composition. Multiple locations separated by several types of delimiters have been found. Spelling mistakes, alternative or historical place names and additional commentary are no exception.

Format	**Amount**	**Percentage**
DD [month] YYYY	25939	62.75%
YYYY	4909	11.88%
[part of] [century]	2345	5.67%
[month] YYYY	1170	2.83%
Easily recognized structures (above)	34363	83.13%
Other recognized structures *	5393	13.05%
Tales with no date value	1580	3.82%
Total tales	41336	100%

** Values containing enough information to compose structured date values*

Table 3: Statistics about the different date formats that were found in the DFDB.

Format	**Amount**	**Percentage**
Place (Province)	32284	78.10 %
Place name only	1466	3.55 %
Province only	1205	2.92 %
Easily recognized structures (above) *	34955	84.56 %
Other recognized structures **	952	2.30 %
Items without geographic information	5429	13.13 %
Total tales	41336	100 %

** Values containing enough information to retrieve additional geographic data*
*** Including all non-Dutch/Belgian locations and exotic formats*

Table 4: Statistics about the different geographical formats that were found in the DFDB.

An alternative way to specify a geographic reference in the DFDB is to assign a *Kloeke georeference*. This is a geographical code for the place where the story

was told. To facilitate his dialect research, in 1926 Gesinus G. Kloeke (1887-1963) divided the map of the Low Countries into a grid and added codes to (most) places. The system was long ago adopted by the Meertens institute as the geographical standard. It will remain to be supported in the future because many books, papers and publications make use of it.

The *source format* data type has always been filled using a selection list and contains abbreviated values, nicely conforming to the standard. It holds a code for the type of source from which the story originates. For instance, B stands for 'boek' (book) while M stands for 'mondeling' (oral), meaning that the story was recorded from oral transmission.

In the fields for *motifs, subgenre*, and *keywords* we see similar problems as for date and geographic location. Several types of delimiters have been used, and various ways to indicate uncertainty about the assigned values. This can make searching and separation of values problematic.

In the *type* field we found 24 values with typing errors that were easily traceable, but also 65 values that could not be found in any of the tale type indexes used by the DFDB. No controlled vocabulary of *keywords* was defined for this collection. A keyword can be a number, name or a word in any time, state, or language. After extraction of all keywords from the database, we ended up with a total of 562480 assigned keywords of which 41555 are unique. Of the assigned keywords, 993 had additional commentary, disclosing information about the context of the keyword, while 50 contained question marks or square brackets to denote uncertainty of meaning or relevance. Some (translated) examples are: 'mirror [black]', '[cannibalism]', 'piggy (?)', and 'punishment?'.

For the fields containing person names, such as *collector* or *narrator*, input conventions have been appointed but these have not always been respected. Most frequently the name is written as 'surname, first name' but often deviations like the reverse, comma-less, title plus name, or just first name have been used. In more recent items, obtained from the Internet, forum user names have been entered. Since it is not always clear which name is the first name, this is a hard problem to tackle fully automatically.

4.2 Standardization of DFDB metadata

It is possible to determine an order of importance in the standardization of the values in a cultural heritage database based on the search and browse behavior of users. The values that are most often used for search, sorting and visualization in the DFDB are all fields, title, keywords, dates and geographical location. Here we discuss the standardization of dates and geographical locations in detail, and treat the rest as standardization problems of a similar nature. With this standardization step we take into account the properties and limitations of the chosen metadata standard, Dublin Core.

Dates As explained above, the date values in the DFDB have always been entered in free text. They can range from a perfectly formatted date value to a complete sentence with comments. To improve functional searching, sorting and conversion, we need to be able to capture all this information in a simple computer readable format. Most free text date values represent either a single date or a time span. Therefore we chose to adopt a data container with a range of two dates defined as 'from', and 'up to and including', both conforming to the ISO 8601 standard. This standard defines a date as YYYY-MM-DD. If an item's date is a single date, both these dates will be the same. For existing free text values, we propose a strictly defined interpretation, determined in consultation with the humanities researchers maintaining the DFDB. This interpretation is shown in Table 5. Strict values have been determined to represent 'end of', and 'beginning of' indicators. The observed 'before', and 'middle' or 'halfway' values have also been quantified.

Description)	**Definition**
midway point century	YY51-01-01
midway point year	YYYY-06-01
midway point month X	YYYY-[X-(MAX_MM_DATE/2)]
beginning of / end of century	first/last 20 years
beginning of / end of year	first/last 2 months
beginning of / end of month	first/last 7 days
mid/halfway century	10 year window around midway point century
mid/halfway year	1 month window around midway point year
mid/halfway month	3 day window around midway point month

Table 5: Quantified definitions of free text dates. MAX_MM_DATE stands for the last day of the month in question

In the future, users will be allowed to enter a date in free text format that is supported by the definitions above. This will then be automatically translated as shown in Table 6. All structured date ranges will be placed in the Dublin Core date field after processing and approval of the annotator. Additional comments concerning a date will have to be specified in the item's comments field. The date range from the item will in turn be used to generate human readable dates like '14th century', or 'winter 2012' for viewing. A somewhat similar approach to accommodating users' preference for 'common language' over strict date formats is that of Petras et al. [7], who map named time periods (e.g., the Renaissance or the Cold War) to date ranges.

Geographical locations For the 'region' field of the narration of a folk tale, the original, and most frequently occurring format is 'place_name (province)'. The value is based on the name of the location at the time of narration. At present, it is hard to search the DFDB for tales that were recorded in a particular county or

Description	Translation
Precise date	(example) 1550-01-01 - 1550-01-01
Cth Century	[(100(C-1))+1]-01-01 - [100C]-12-31 (official)
Only YYYY available	YYYY-01-01 - YYYY-12-31
Only YYYY-MM available	YYYY-MM-01 - YYYY-MM-[MAX_MM_DATE]
Xth quarter of year YYYY	YYYY-[3X-2]-01 - YYYY-[3X]-[MAX_MM_DATE]
Xth quarter of Cth century	[(100(C-1))+1+(25X-25)]-01-01 - [(100C)+(25X)]-12-31
Beginning of Cth century	[(100(C-1)+1)]-01-01 - [(100(C-1)+20)]-12-31
Beginning of year YYYY	YYYY-01-01 - YYYY-02-31
Beginning of month MM	YYYY-MM-01 - YYYY-MM-07
End of Cth century	[(100(C-1)+81)]-01-01 - [(100(C))]-12-31
End of year YYYY	YYYY-11-01 - YYYY-12-31
End of month MM	YYYY-MM-[MAX_MM_DATE-7] - YYYY-MM-[MAX_MM_DATE]
[season] year	[Begin date season] - [end date season] (Outer possible dates of that season)

Table 6: Strictly defined interpretations of free text date values

region. When taking future international cooperation into account, it is preferable to supply higher order information such as country and continent name. A suitable hierarchical order for locations would be: Geographical coordinate (latitude, longitude), spot (building name/artwork/tree/dune), street (with optional number), place (village/city/lake), county, region (area/nature reserve/mountain), province, country, and continent (and perhaps even planetary body for future entries). Some examples:

- Full set: (53.360304, 5.214203), Brandaris, Torenstraat, West-Terschelling, Terschelling, - , Friesland, the Netherlands, Europe, Earth
- Partial set: (52.37403, 4.88969), - , - , Amsterdam, Amsterdam, - , Noord-Holland , the Netherlands, Europe, Earth
- Partial set: (51.74308, 4.77339), - , - , - , - , Biesbosch, Noord-Brabant, the Netherlands, Europe, Earth

An open source geographical database that can supply such information is Geonames.[6] This database contains roughly 8 million geographical entries worldwide, and their corresponding coordinates, in a hierarchical manner. The items already present in the DFDB will be supplemented with all available information that can be retrieved from Geonames. For future input however, Google Places[7]

[6] http://www.geonames.org/

[7] https://developers.google.com/places/

will be used for retrieval of coordinates and hierarchical geographic information. We will do this using a simple geolocation plugin for the Omeka content management system which has been extended to fit the information needs of the DFDB.

To facilitate historical annotation of a folktale, we either need to create the freedom for an annotator to supply historically sound information, or to consult a very complete spatial history database. The latter option is being investigated by several groups around the world [8, 9], yet no completely open source solutions are available at the moment. When these become available, it is still possible to adopt them. We leave room in our system for manual alterations in the data provided by Google, so that names of old towns and counties can be supplied.

We will continue to support the Kloeke georeference codes. This way, we assure a link with the other collections of the institute and are still able to use the old visualization methods used by some researchers. It will however no longer be necessary to look up this code manually, adding to the reduction of input steps for a new item. The Kloeke georeference data will not be stored in a standard Dublin Core field since it is useless for researchers outside the Meertens institute.

Tale types, motifs and keywords Since these values have always been typed, or copy pasted into place, some input errors have been made. However, only little inconsistency was observed in these values. Datasets are available for the tale types and motifs that were cross referenced for correctness. This yielded lists that were of manageable size for manual curation.

The generated list of unique keywords will be used to give suggestions by means of auto-completion when an annotator attempts to add them. This will prevent further addition of variants. In a sense, this is an ideal compromise between a controlled vocabulary and complete freedom.

4.3 Meta-metadata standardization

If an annotator is uncertain about a metadata value, it should be somehow stored in the data set. In case of automatic annotation it is desirable to assign a confidence level to the annotation. An annotator can consequently approve or reject the outcome. This type of information can be considered meta-metadata. It is currently specified for the following DFDB fields: Tale type, Narrator, Kloeke georeference (location of narration), Kloeke georeference (locations in the tale).

Meta-metadata could be stored in the following ways:

- A special character or note in the data itself, e.g. #D:[confidence level]
- An additional term or column in the database, e.g. location_disputed (yes/no)
- An additional dataset containing the value's id and status, e.g. value_id, status, status_score, date_created

A *special character* in the data itself might confuse the user of the data, and can be seen as data pollution. An *additional term* in the database is a good solution

for a “quick fix". The disadvantage is that for every metadata descriptive term a column has to be added to the database. The last option from the list is the most permanent and modifiable solution. An additional system can monitor the state of each value of each item of the database. An additional table can be created containing the item's value's id, a meta-metadata type, a score and a date. Now, an item can be flagged as disputed, or as being computer generated with a confidence interval of 0.43, or approved by a user. Table 7 shows some examples of these values.

item_value_id	meta-metadata-type	confidence	date_created
(item 1, tale type)	disputed	-	01-01-2001
(item 25, language)	generated	0.43	10-01-2010
(item 25, language)	approved	-	11-01-2010

Table 7: A few examples of rows in a meta-metadata table

4.4 Challenges

The DFDB, like many other online cultural heritage collections, is not static but is constantly being supplied with new items. Because annotators and users rely on a working system, everything should stay that way. The standardization process has to be carefully planned and prepared; trial and error will confuse the users and cause even more errors and inconsistencies. The process is best carried out in big steps. It can be compared to repairing a moving bicycle.

5 Conclusion

At first glance, the standardization of a cultural heritage database seems like a straightforward job which could be done through some simple filter actions, value separations, and regular expressions. However, the contrary is true. Many factors have to be taken into account before the actual data can be touched and changed. The data need to be studied in detail to get an overview of all the possible types of values. Standardization of data brings along a large set of problems. We have shown that manual annotation of cultural heritage databases can spawn several types of errors and inconsistencies. We have discussed the choice of metadata standards and infrastructures and how to connect with them. Existing data standards and conventions will have to be respected, or developed when none are present. The longer conventions are not followed, the larger the divergence in values becomes, and the harder the clean up operation will become. On top of this, the addition of items to the database by annotators is a continuous process which makes a faceted standardization of the database and its functions a difficult task. We hope that this paper can be a helpful asset for future endeavors of a similar kind.

6 Acknowledgments

This work has been carried out within the Folktales as Classifiable Texts (FACT) project, part of the Continuous Access To Cultural Heritage (CATCH) program funded by the Netherlands Organization for Scientific Research (NWO).

References

[1] James Abello, Peter Broadwell and Timothy R. Tangherlini. Computational Folkloristics, *Communications of the ACM* 55(7): 60–70, 2012.

[2] Dong Nguyen, Dolf Trieschnigg, Theo Meder, and Mariët Theune. Automatic Classification of Folk Narrative Genres, *Proceedings of the Workshop on Language Technology for Historical Text(s) (KONVENS 2012)*, pp. 378–382, 2012.

[3] Dolf Trieschnigg, Djoerd Hiemstra, Mariët Theune, Francisca de Jong, and Theo Meder. An Exploration of Language Identification Techniques for the Dutch Folktale Database, *Proceedings of the Workshop on Adaptation of Language Resources and Tools for Processing Cultural Heritage (LREC 2012)*, pp. 47–51, 2012.

[4] Antal van den Bosch, Marieke van Erp, and Caroline Sporleder. Making a Clean Sweep of Cultural Heritage, *IEEE Intelligent Systems* 24(2):54–63, 2009.

[5] Theo Meder. From a Dutch Folktale Database towards an International Folktale Database, *Fabula* 51(1-2): 6–22, 2010.

[6] Tamás Váradi, Steven Krauwer, Peter Wittenburg, Martin Wynne and Kimmo Koskenniemi. CLARIN: Common Language Resources and Technology Infrastructure, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, 2008.

[7] Vivien Petras, Ray R. Larson, Michael Buckland. Time Period Directories: a Metadata Infrastructure for Placing Events in Temporal and Geographic Context, *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, pp. 151–160, 2006.

[8] Richard White. What is Spatial History? Spatial History Lab: Working paper [online] `http://www.stanford.edu/group/spatialhistory/cgi-bin/site/pub.php?id=29` [Last accessed 2012-09-17]

[9] Kilian Schultes and Stefan Geißler. Orbis Latinus Online (OLO), Digital Humanities 2012, July 2012. `http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/orbis-latinus-online-olo/` [Last accessed 2012-09-16]