The temporal evolution of the normalized web distance

Is a "Wirikuta empowerment" of the Huichol measurable on the internet?

Lorena Perez-Garcia, Jan Broekaert and Nicole Note
Center Leo Apostel for Interdisciplinary Studies, Vrije Universiteit Brussel, Brussels, Belgium

Received 22 July 2014
Revised 8 October 2014
5 May 2015
10 July 2015
27 November 2015
11 December 2015
Accepted 14 December 2015

Abstract

Purpose – The purpose of this paper is to assess whether the temporal evolution of the normalized web distance (*NWD*) between significant terms concerning, e.g., a case of online activism can be used as a meta-data technique to measure evolution over time of, e.g., progress or decline of social empowerment.
Design/methodology/approach – The *NWD* between two terms has been identified as a quantitative measure for semantic proximity, ascertaining a defining relation between them. A trend analysis is made by performing on the internet a time window restrained series measurement of *NWD* of all combinations of key-terms and classifier-terms. Case defining key-terms, positive and negative discourse polarizing classifier-terms, and neutral classifier-terms for negative control need to be determined by discourse analysis of information on a targeted case. An example of *NWD* evolution from 1994 until 2013 is presented to measure the empowerment effects of the Wirikuta online movement on the Huichol people in Mexico.
Findings – The application of the *NWD* temporal evolution method to the Wirikuta case shows a slight but significant semantic change of the key-terms with respect to some of the positive and negative classifier-terms. The neutral classifier correctly shows no significant distance variation, as required for valid application of the method. The method provides indications for a complex image of empowerment of the Huichol identity.
Research limitations/implications – The accuracy of the method is limited due to short-term and between-user variability of the search tool's page counts. More reliable access to a web-index will be required for more accurate *NWD*-based trend analysis.
Practical implications – The monitoring of temporal *NWD* evolution provides a potential tool for more comprehensive trend description compared to classical frequency based methods.
Originality/value – Trend analysis is key to internet research, to which the temporal *NWD* method provides an innovative contribution.

Keywords ICT, Normalized web distance, Social empowerment, Temporal evolution
Paper type Research paper

Introduction

The internet entices millions of users around the globe to gain access to billions of web pages and social media entries, and actively create online content.

In this way the internet constitutes a self-organizing virtual and heterogeneous public sphere (Rheingold, 1993; Castells, 2008; Freyermuth, 2010). Indeed, technologically and culturally instant communities of transformative practice emerge within this public space (Castells, 2012). During the past decade, the appearance of social media platforms and the increase in their usage changed not only technological – consumer – trends but also communication and awareness in society at large. Close to literally, "the world" witnessed the uprisings in Tunisia and Egypt in 2010-2011 and with them the rise of empowered

Authors Lorena Perez-Garcia and Jan Broekaert contributed equally to this work.

Internet Research
Vol. 26 No. 5, 2016
pp. 1269-1290

1066-2243
DOI 10.1108/IntR-07-2014-0185

sectors of society through the use of social media. Freedom of expression on the internet and personal spreading of messages through "mass self-communication" media became a new asset to society (Castells, 2012). Grassroots social movements found in social media the opportunity to have an impact on individuals and society, without having to rely on either physical space or real-time contact.

The individual agents in this process co-create the dynamics of a social movement and potentially empower social – but not necessarily local in a physical sense – communities.

The complex entanglement of agonist and antagonist elements of online information, the wide spectrum and variety of interests, views and ideologies leads to an imbroglio of information which permanently changes over time. As a result, the majority of online resources is often only meaningful in a low-quality approximate sense (Oliver *et al.*, 1997). Given the vast amount of information on the internet on a given subject, beyond the scope of individual perusal, could one still obtain meaningful information from meta-data about them? That is, from the number of specific query-relevant documents and the time of publication of query-relevant documents, instead of the information content in them? Would it be possible to measure trends and evolutions of an empowerment – in particular its reflection in the public sphere of the internet – from a meta-data perspective? (Schoen *et al.*, 2013).

We encountered this research problem when we were evaluating empowerment through ICT in indigenous communities (e.g. Joseph, 2013), in Mexico (Pérez García, 2016). We hypothesized that the empowerment of "local" use of ICT could be reflected in its global perception, forwarded by the internet. In particular we applied our temporal *NWD*-method to a case in which a local, diasporic and sympathizing global "community" acts to defend the Wirikuta, the Huichol's sacred land, from open-surface mining projects by international contractors (see Appendix 2).

Our central research goal was to develop a method which could probe in a meta-level manner the internet for empowerment achieved in communities. This was not based on the perusal of individual online documents but by an extensive evaluation of "bulk" online content. This hypothesis supposes a strong relation between the deploying of relevant events and its "same time" coverage on the internet (see e.g. Jungherr and Jürgens, 2013).

In Appendix 2 we illustrate how the various stages of the Wirikuta case have been documented on the internet during, or following closely, their deployment. The data content of the internet should thus reflect the evolution of the events, in as much as the internet retains its history. Since providing online information is never done unintentionally, the relation between online data and offline events needs to be approached critically (Jungherr and Jürgens, 2013). Demographic selectivity toward internet usage in general and personal and institutional objectives steer the content of online information. Also, the "unvoiced" are not expressed on the internet but may be paramount in situational real world events. In many cases new content is written over older content and as such this older information would be lost. Just as frequently, however, the older information is reproduced at other web locations before it is replaced at its original location and thus not updated, time-tagged information remains available on the internet. Issues like link rot may indicate information is either no longer available at the link, or possibly relocated to another URL. But also straight removal of information – sometimes as a result of the "right to be forgotten" on privacy grounds (European Commission, 2012) – occurs, while at the same time active historical archiving of the internet has incited "the Net never forgets" (Lasica, 1998). While to a certain extent these effects alter over time the historic on/off line mirror representation,

we reckon the relatively small number of irretrievable documents will only affect the *NWD*-method negligibly.

In the next section we will develop a "tree-ring-growth structure" of the internet and investigate how this model enables a temporal meta-data analysis of our example case concerning Wirikuta.

Quantitative monitoring of the internet and social media has been developed in a variety of ways, each targeting specific objectives (Schoen *et al.*, 2013).

In relation to our present method we overview some recent developments that surpass mere frequency monitoring for trend analysis:

- The drop of diversity of hashtags on Twitter has been proposed as an indicator of unexpected or extraordinary real world events (Jungherr and Jürgens, 2013). Temporal analysis of the Shannon diversity index in top-1000 hashtags frequencies into diurnal rhythm, trend and remainder has shown a concentration of user interest during extraordinary offline events.
- A semi-automatized method for predicting credibility of tweets related to news events was devised by Castillo *et al.* (2013). This method initially relies on manually tagged data of peak news events from Twitter Monitor, and after machine learning the process provides an automatized feature related (based on message, user, topic and propagation properties) credibility score.
- The common rationale of predictive modeling of social media was outlined by Kalampokis *et al.* (2013); a critical approach to data-conditioning followed by specific forecasting methods for time series. The latter clearly targets anticipation of new observations rather than explanation of data correlation to events.

Our present exploratory study of temporal evolution of the *NWD* differs by focussing on long-term evolution as compared to data peak-identified events of extraordinary nature. It is not our focus to forecast, but rather to correlate the evolution with the unfolding of events in our case study. Our targeted concept of "identity empowerment" (Navarrete Linares, 2009) is not an ephemeral phenomenon nor a simply quantifiable parameter, much like the concept of "user sentiment" in Twitter-related predictive methods. We approach empowerment by the shifting of semantic clusters of key-terms describing socio-cultural classifiers relevant to our case.

To establish our study we (Section Normalized Web Distance (*NWD*)) introduce the *NWD* and extend its use to describe temporal evolution of semantic similarity (Cilibrasi and Vitányi, 2007); (Section Methodology) describe the methodology of the internet experiment; (Section Data) assess the hit-counting using the Google search engine and the accuracy of the temporal *NWD* evolution method; (Section Conclusion) conclude our study cautiously affirmative on the relation of *NWD* evolution and empowerment in the Wirikuta case. Appendix 1 provides all *NWD* temporal evolution graphs and Appendix 2 describes stages of the case of the Wirikuta offline and online movement. The collected web count data for this study are available from the authors by e-mail request.

The *NWD* and its temporal evolution

NWD

Web search engines have become valuable tools to research a subject in the electronic, global public sphere. A simplified quantitative approach to evaluate the extent of a case – not necessarily its presence and priority in the classical online news channels – would be the absolute number of related documents on the internet in general. This could be

revealed by counting pages with case-relevant terms via web search engine. This absolute number would have some significance for the prevalence of the case, given the appropriateness of the query terms. However, the prompt hit count does not provide insight in the interpretation of this case nor in its unfolding over time. Furthermore, each hit count does not necessarily coincide with a unique document, since content is often integrally or partially copied and redistributed to other URLs.

Instead of using some absolute hit count as an approximate online proxy indicator of corresponding offline activity, we propose the *NWD* (previously less neutrally denoted as "Normalized Google Distance") for its capacity to measure semantic relations. We use semantics here as in the operational approach of Cilibrasi and Vitányi (2010) – "The semantics of a word or phrase consists of the set of web pages returned by the query".

A small *NWD* between two terms does not mean they are semantically interchangeable in a sentence, it indicates the terms are in a meaning cluster of close relationship in natural language; a relation which is either feature defining (e.g. blue and sky), syntactically alike (e.g. *the* and *and*) or categorical inclusive (e.g. vegetable and carrot). For example, the terms the and and have very high co-occurrence on the internet and thus have strong semantic similarity, so strong their *NWD*(the, and) is approximately 0 (Figure A2). This definition of internet semantics, however, also leads to the effect of small *NWD* between semantically opposite terms; namely for qualitative variations like black and white.

It must be noted that next to co-occurrence of the terms an increasing number of documents on the internet in general will lower the *NWD* (see Equation (1)). The latter could be interpreted as an emphasis of the relation of semantic similarity between two terms when their co-occurrence appears more insular in the larger totality of web documents.

The *NWD* does provide the tool for a possible online proxy indicator for the semantic defining of some offline event: it allows to quantitatively assess the connotation of its main descriptors. Suppose a case of interest would have the name of an indigenous community as a key-term for its description, e.g., Huichol. Then it would be of interpretative significance to know the connotations of this term. Is the key descriptor Huichol semantically closer related to, for example, violence than it is to sacred land? Even more interpretatively valuable is the answer to the question: how does the connotation of these key descriptors evolve over time during which the case of interest is developing? Well-chosen *NWD* could thus provide an online proxy indicator for the perception of offline events. In particular a number of discourse polarizing classifier-terms should be assigned around the key descriptors, such that the evolution of the *NWD* of these "positive" and "negative" classifiers with regard to the central terms provide information for the interpretation of the case's evolution. Then these *NWD* of well-chosen key descriptors of, for example, a community as a proxy indicator can provide an insight into the struggle for their causes and thus an insight into their empowerment. In the application example of our *NWD*-based method we precisely envisage the possibility to measure by proxy this empowerment.

Some caveats concerning the *NWD* interpretation need to be taken into account. Since the *NWD* method only uses meta-data for the relation between semantics and meaningful information, a number of artifacts may occur. The latter relation is not only subject to interpretation but also to effects of syntax, polysemy, metonymy, and language as discourse itself. The sole measurement of term co-occurrence cannot avoid these effects of natural language. Some of these effects can be attenuated by tweaking

selected key-terms and classifier-terms. Pure semantic relationality is not unambiguous for meaning interpretation the syntax of a sentence in which a term occurs will to large extent determine its connotative meaning, e.g., "Michael Jackson hits" vs "Michael hits Jackson"; polysemy of terms, e.g., fashion will mean both "a manner of doing something" or "a trend of style"; metonymic use of terms like *s*acred land or music in natural language, for example, an indisputable conviction or a much desired answer; and language mixing – Wixarika terms are lent to Spanish and lent to English which leads to measuring *NWD* using terms of a different languages. Since the perspective of the internet is global, the use of English terms as its *lingua franca* was acceptable in our approach to the Wirikuta case.

Other effects on search-based methods include the possible issue of measuring the internet in a "filter bubble" (Pariser, 2011), which distorts the measurements since only a selective segment of the internet is probed. Thus procedures of commercial optimization and user predilection adaptation by internet search engines, and institutional censorship could not only modify the ranking but also the quantity of query results. Finally, we remark that the *NWD* expression Equation (1) is not sensitive to result ranking nor to scaling (equal relative change of M_y, m_y, μ_y and N_y leave the *NWD* invariant). Moreover, the sheer size of the internet and the asynchronous global distribution of the web-indexing data causes search tools to return estimated counts instead of true numbers of pages (Satoh and Yamana, 2012), according to commercially protected algorithms.

Temporal evolution

A prompt evaluation on the internet of case relevant *NWD* between key-terms and key-classifiers would return a measure, that is composed of all online information uploaded over time. More informative would be to know the evolution of connotations of the key descriptors over time. A longitudinal study would provide some insight but requires an equally long term monitoring of the internet. One, however, can at a given moment reconstruct the time evolution in retrospect by applying time window restricted queries to make a historical reconstruction of the connotation of a term. Only documents uploaded within the time window – with its prevalent semantic connotations – are searched. In order to have a meaningful interpretation of these temporal *NWD*-variations a "negative control" measurement needs to be available as well. This measurement consists of providing a priori a well-chosen term which from interpretative perusal of documents would appear indifferent to the key-descriptors of the case. In order for this negative control measurement to be meaningful a term should be selected which is not overly rare. In our example case of the Wirikuta one neutral term was retained after perusal of sampled documents, namely, table. An additional verification of our measurement of the temporal evolution of the *NWD* consisted in the measurement of two common English terms *the* and *and*, which we consider to show an immutable standard relation of the English language.

Since the *NWD* also requires to fix the size of the sampled corpus, we have chosen to track the increment of the page count of the term the, multiplied by a constant factor equal to 100. This procedure consistently keeps track over time of the relative size of each increment or "tree-ring." The chosen factor 100 was set to keep all *NWD* within the range between 0 and 1 (Cilibrasi and Vitányi, 2007).

The sample frequency of the *NWD* was adjusted to the envisaged purpose of tracking proxy changes of empowerment. This focus would not be reflected by changes due to peak offline events, but rather follow a slow modification over time by seeding of

global cultural apprehension. We considered that the modest extent of our example case on the Wirikuta (compared to the Egyptian uprising) with restricted local ICT, would not systematically lead to a fast, nor peaked, extensive expression on the internet. Therefore the time window set to one year will average out ephemeral short-term variations of *NWD* and expose durable change of semantic relations. The temporal evolution for an annual sample frequency is obtained by restricting for each year Y the time window of the query to 01/01/Y – 12/31/Y, and recording the number of page hits for each query. Each query is thus time-restricted to the domain of an annual increment of the internet, corresponding to probing consecutive "tree-rings" of the growing internet – again with the proviso that the loss of information on the internet over time is limited. The validity of this approach requires a strong time-tagging of a document which should be correctly identified by the search tool (e.g. Webster, 2007, for a study using Google Scholar Search).

Technically the *NWD* between two terms provides a measure between 0 and 1 that weighs their co-occurrence, single occurrences and internet size according to the amount of Kolmogorovian information necessary to transform the first term into the second (Cilibrasi and Vitányi, 2007). The necessary meta-data for the temporal *NWD* measurement are the number of single query returned documents; the time of publication of query-relevant documents; the number of joined query returned documents for co-occurrence; and the approximate annual increment of the internet. For two terms u and v, with respective hit-counts $n_y(u)$ and $n_y(v)$ and with $\mu_y(u \text{ AND } v)$ the count for the search "u AND v", each in the temporal confinement of the year y, the $NWD_y(u, v)$ is given by:

$$NWD_y(u,v) = \frac{\ln M_y - \ln \mu_y}{\ln N_y - \ln m_y} \tag{1}$$

where $M_y = \max\{n_y(u), n_y(v)\}$, $m_y = \min\{n_y(u), n_y(v)\}$ and N_y is the increment of the internet in the year y. The latter is a number – following Cilibrasi and Vitányi[1] – related to the annual increment, that is chosen in order to keep the maximum *NWD* below 1. We have chosen to use as the annual increment of the page count of the common term *the*, multiplied by a constant factor set to 100[2].

In the tree-ring growth-model of the internet, we assume that in each consecutive layer the relative amount of n_y(u), n_y(v), μ_y(u AND v) and N_y vary (Figure 1).

The temporal variation of the *NWD* can be broken down accordingly:

$$\delta NWD(u,v) = \frac{1}{\ln N - \ln m}\left(\frac{\delta M}{M} - \frac{\delta \mu}{\mu} + NWD(u,v)\left(\frac{\delta m}{m} - \frac{\delta N}{N}\right)\right) \tag{2}$$

with $\delta M = M_{y+1} - M_y$, idem for the other increments. In the event of an observed change of *NWD*, it is possible to inspect its precise origins in $\{\delta M, \delta m, \delta \mu, \delta N\}$.

Methodology

In our Wirikuta case study of the temporal *NWD* method three key-terms were selected to identify our case of focus (see Appendix 2). Also, 13 classifier-terms were chosen and subdivided into three categories; positive, negative and neutral, according the alleged type of discourse they are related to. The list of positive classifiers includes terms that would normally be related positively to culture, hence determined as indicative – or with importance – for the identity of the concerned people; the Huichol. The list of the negative classifiers in the Wirikuta case contains terms that could be perceived to have a negative

relation to the Huichol culture. For the three key-terms central to the "Wirikuta case" we have chosen Wirikuta, Huichol and Wixarika. These key-teems unequivocally identify the core elements of the case; the Huichol indigenous community and their target of activism, the Wirikuta. Since the Huichol name themselves in their proper language the Wixarika, the latter term is considered key as well. The seven selected positive classifiers are; sacred land, Marakame, peyote, ancestors, Wixaritari, music, fashion. The five selected negative classifiers are; violence, addiction, discrimination, racism and mines. The choice of these classifier-terms is to certain extent underdetermined; the terms have been selected after analyzing the various perspectives which cross the Wirikuta case (see also Appendix 2), such as economic benefits, cultural precarity, and discriminatory prejudices. These selected classifier-terms have been chosen to provide insight in possible polarizing discourses concerning empowerment, it is however possible to determine other classifier-terms in order to gain insight on other aspects of the online evolution of the Huichol. In order to have a negative control to our interpretation of approaching or receding *NWD* over time, we considered a single term that would maintain an invariable semantic relation with respect to the Huichol culture, and chose the term table[3].

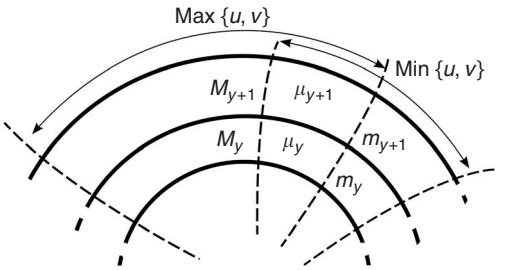
Finally a measurement of the temporal evolution of the *NWD* of two common English terms – *the* and *and* – were used as well in order to control the negative control based on table.

The measurements consist of manually executing queries on a freely available search engine – Google Search – which provides an estimate of the hit counts and which allows a "Custom date range" functionality to a query. The date range was chosen to coincide with the calendar year 01/01/Y – 31/12/Y. Starting with Y = 1994 up till $Y = 2013$ the hit counts of all queries were recorded.

In order to calculate the $NWD(u,v)$ also joint queries of key-terms u with classifier-terms v were made. The query terms always were embedded in quotation marks – "term" – in order to avoid stemming or associative query results which tend to be provided by search engines. For joint queries the Boolean operator AND was added explicitly in the query: "*u* AND *v*". In addition the key-terms were also checked for *NWD* temporal variation.

Considering the variability of Google's estimation algorithm for hit counts, we worked on three different computers, from different IP addresses. The hit counts for the queries were retrieved in the period February 17, 2014-February 24, 2014, in the Brussels region in Belgium[4].

In order to estimate these errors, queries were repeated on three different computers at a different moment. For single queries five repetitions were carried out ($N_{\exp} = 5$).



Notes: Relative changes M_y, m_y, m_y and N_y in comparison to previous year give rise to change of the $NWD(u, v)$

Figure 1.
Scheme of tree-ring growth-model with max, min and joint counts indicated for year *y* and subsequent year *y*+1

while joint queries were repeated three times ($N_{exp} = 3$). The standard error on the averages was derived from the standard deviation of the gathered hit counts. A priori the Google Search tool seems to set the error on hit numbers by providing three significant numbers. This sometimes leads to quixotic outcomes in the lower range spectrum. Given the observed between – user variability we consider this a vast underestimation of the accuracy (Kilgarriff, 2007).

Data reduction and interpretation

Volume growth

The growth of the internet is clearly apparent from the "date range" queries; the tree-ring growth shows an exponential increase after 2001 (near linear trend in logarithmic scaled Figure A1).

From our measurements of absolute counts we notice exponential growth of Wirikuta, Wixaritari, Wixarika and Marakame over the full time period of our measurement 1994-2013: all of them are rare Wixarika language terms that start at zero or almost zero counts. The term Huichol – closely related to Wixarika – starts at higher counts in 1994 and does not expose this exponential growth until 2001, as all terms with higher initial counts do. Given the possibility of artefacts in the count estimate algorithm of Google Search it is not possible to confer any interpretation to this difference in growth prior to 2001[5]. These artefacts disappear starting 2001 and led us to discard any interpretation of data with counts higher than approximately 100 in the period 1994-2000.

Of all 18 terms measured – key, classifier, control – the two fastest growing terms were Wixaritari with slope 0.21 ± 0.02 ($R^2 = 0.91$) and Wirikuta with slope 0.20 ± 0.02 ($R^2 = 0.94$). The slowest growing term was table, with slope 0.10 ± 0.01($R^2 = 0.90$).

The *NWD* requires counts for joint queries using the Boolean connector AND. However on examination Google Search returns hit counts that do not accurately satisfy the inclusion-exclusion principle (the OR-measurement does not equal the sum of the separate measurements minus the AND-measurement). Therefore one should critically assess the Boolean conjunction in queries.

Semantic clustering

First we observe the *NWD* adequately expresses semantic similarity in the sense that small distances are found for closely related terms:

- For Wirikuta (Figure A3, $Y = 2013$) we notice the nearness in increasing order of the terms: Wixaritari (0.10 ± 0.07), Wixarika (0.19 ± 0.04), Huichol (0.20 ± 0.09), Marakame (0.23 ± 0.07), sacred land (0.29 ± 0.08) and peyote (0.31 ± 0.10).

  Larger distances are noticed for mines (0.62 ± 0.07), ancestors (0.63 ± 0.09), and all other negative classifiers around 0.70-0.75 as well as music and fashion 0.75-0.80 and the control classifier table (0.77 ± 0.08).

- For Huichol (Figure A4, $Y = 2013$) we notice the nearness of the same terms: Wixaritari (0.23 ± 0.09), Wixarika (0.23 ± 0.04), Wirikuta (0.20 ± 0.09), Marakame (0.25 ± 0.07), sacred land (0.36 ± 0.09) and peyote (0.26 ± 0.11).

  More remote terms contain again the negative classifiers (range 0.65-0.75) – mines (0.66 ± 0.10) – music (0.75 ± 0.10), fashion (0.68 ± 0.09) and the control term (0.66 ± 0.09).

- For Wixarika (Figure A5, $Y = 2013$) we notice again the nearness of the same terms: Wixaritari (0.07 ± 0.08), Wirikuta (0.19 ± 0.04), Huichol (0.23 ± 0.04), Marakame (0.15 ± 0.07), sacred land (0.33 ± 0.09) and peyote (0.32 ± 0.10).

  More remote terms again contain the same negative classifiers (range 0.70-0.75) – mines (0.65 ± 0.09) and, music (0.80 ± 0.09), fashion (0.71 ± 0.11) and the control term (0.78 ± 0.10).

Closest *NWD* are thus found among terms of the same language Wirikuta, Wixarika and Wixaritari. The term sacred land is marginally more related to the land Wirikuta and Marakame is closer to the people named in their own language, Wixarika, than in Spanish, Huichol. Notice here that mines has about the same distance to all three key terms (measurement 2013).

Semantic evolution

The evolution of the negative control $\langle NWD_y(\text{table})\rangle$, where the average has been taken over the key-terms, and its additional proper control NWD_y(the, and), show a constant value starting in 2002 (Figure A2). The linear approximation of the negative control *NWD* evolution of table with respect to Wirikuta has a slope 0.001 ± 0.001 ($R^2 = 0.148$), for Huichol slope -0.0036 ± 0.001 ($R^2 = 0.579$) and for Wixarika slope -0.001 ± 0.001 ($R^2 = 0.072$); on average we find for the *NWD* of table to the key terms the slope is -0.001 ± 0.001 ($R^2 = 0.153$). The supplementary control assessment by use of the invariable semantic relation of *the* and *and* by *NWD* (the, and) with slope -0.0001 ± 0.0002($R^2 = 0.0057$) shows therefore strong alignment with NWD_y(table) which has a slightly larger variability[6].We conclude that our neutral term table thus indeed exposes no significant change of semantic similarity with the key terms, as we had hypothesized. This entitles – within set limits of error – us to meaningfully interpret *NWD* temporal evolutions of terms after 2001 using this method. However, given the standard error due to the variability of the counts we will adopt the rule that after linear regression at least a slope of ± 0.005 is necessary to hint at a significant trend of change of semantic similarity, while the relative error should be not more than one-third of the slope value. (The interval of 12 years then leads to an approximate minimal change of 0.06 in NWD.) Evolutions with a lesser slope will be considered constant over time at present annual sample frequency. First we notice in general a lesser variability among *NWD* of Wirikuta (Figure A3) and Huichol (Figure A4) as compared to Wixarika (Figure A5). The lesser count rates for the latter term leads to larger fluctuations of the *NWD*. The observed rather systematic plunge of the *NWD* evolution from the start of our measurements in 1994 to 2001-2002 in the three graphs (Figures A3-A5) can be retraced to anomalous slow growth of the δM in Equation (2). This effect can be explained by the artefactual count estimates returned by Google Search prior to 2001. We will therefore only assess changes of *NWD* starting in 2001. Given the error limits, we enumerate the significant *NWD* evolutions:

- For Wirikuta (data from Figure A3) we observe the gradients of linear regression of significant *NWD* changes (Table I).
- For Huichol (data from Figure A4) we observe the gradients of linear regression of significant *NWD* changes (Table II).
- For Wixarika (data from Figure A5) we observe the gradients of linear regression of significant *NWD* changes (Table III).

Conclusion

First we observed the well established fact that the *NWD* adequately expresses semantic similarity of terms, i.e., a smaller *NWD* distance corresponds to a close relationship of terms in a natural language. The key terms Wirikuta, Wixarika and Huichol are close among them, as well as to their most typical cultural classifiers marakame, sacred land and peyote – among our selected classifier-terms. Larger distances to the key-terms are observed for mines and ancestors, and all other negative classifiers but also for music and fashion and the control classifier table. Closest *NWD* are in particular found among terms of the same language – Wirikuta, Wixarika and Wixaritari. The term sacred land is marginally closer related to the land Wirikuta and, Marakame is closer to the people named in their own language, Wixarika, than in Spanish, Huichol. The term mines has about the same distance to all three key terms (measurement $Y = 2013$).

Crucial to our temporal model is the image that emerges from the trend analysis of *NWD* evolution. We found that the negative control term table effectively remained semantically stable with respect to the chosen key-descriptors. This condition, together with an additional check of the semantic stability between the terms *the* and *and*, provides the ground to allow a significant interpretation of the *NWD*-evolutions.

However, the size of the standard error due to the variability of the repeated counts requires discarding variations in linear regression below ± 0.005 per annum, and with relative error larger than one-third of the slope value. We find a weak indication of semantic change and overall a complex evolution for the classifier-terms with respect to the key-terms (see Table IV). Within our set of classifiers the territory Wirikuta has slightly gained relation

Table I.
Significant average annual changes of NWD (u, Wirikuta), period 2001-2013

Wirikuta	
Receding *NWD*	Approaching *NWD*
Ancestors 0.010 ± 0.002 ($R^2 = 0.616$)	Wixaritari -0.008 ± 0.002 ($R^2 = 0.650$)
Marakame 0.006 ± 0.002 ($R^2 = 0.532$)	Mines -0.005 ± 0.001 ($R^2 = 0.643$)
Music 0.005 ± 0.001 ($R^2 = 0.841$)	

Table II.
Significant average annual changes of NWD (u, Huichol), period 2001-2013

Huichol	
Receding *NWD*	Approaching *NWD*
Racism 0.0130 ± 0.001 ($R^2 = 0.933$)	(none)
Addiction $0{,}009 \pm 0{,}001$($R^2 = 0.899$)	
Music 0.007 ± 0.001 ($R^2 = 0.851$)	
Violence 0.006 ± 0.001 ($R^2 = 0.835$)	
Discrimination 0.005 ± 0.001($R^2 = 0.873$)	

Table III.
Significant average annual changes of NWD (u, Wixarika), period 2001-2013

Wixarika	
Receding *NWD*	Approaching *NWD*
Addiction 0.010 ± 0.002 ($R^2 = 0.723$)	Discrimination -0.016 ± 0.005 ($R^2 = 0.591$)
	Sacred land -0.012 ± 0.003 ($R^2 = 0.570$)
	Fashion -0.007 ± 0.002 ($R^2 = 0.489$)
	Mines -0.005 ± 0.001 ($R^2 = 0.578$)

	Wirikuta	Huichol	Wixarika	Table	Sacred land	Marakame	Peyote	Ancestors	Wixaritari	Music	Fashion	Violence	Addiction	Discrimination	Racism	Mines
Wixarika	X	0	0	0	0	+	0	+	–	+	0	0	0	0	0	–
Huichol	0	X	0	0	0	0	0	0	0	+	0	+	+	+	+	0
Wirikuta	0	0	X	0	–	0	0	0	0	0	–	0	+	–	0	–

Notes: Semantically approaching terms or diminishing *NWD* indicated by "–". Semantically receding terms or increasing *NWD* indicated by "+". Absence of significant change of *NWD* between terms is indicated by "0"

Table IV. Qualitative indication of significant semantic change between key terms and classifiers over the period 2001-2013

with its proper named people Wixaritari and its disputed mines and slightly lost relation with the English term ancestors and the name for the shaman, Marakame. The term Huichol has taken slightly more distance from most of our selection of negative classifiers, but also from music. Finally, the adjective and language name Wixarika has a closer relation to the negative classifier-terms discrimination and mines but also to fashion and sacred land, and had slightly less relation to the term addiction. To sum up, we conclude that a redefinition of the cultural homeland Wirikuta has included the struggle of the Huichol community with the consequences of mining. The online data reflect an evolution of less association of Huichol with negative connotations – as taken into account by selected classifier-terms in our experiment. From the perspective of the evolution of semantic clusters we can thus affirm a slight identity empowerment of the Huichol. Nevertheless, we recall the ambiguous relation between offline transformations and changing online data remains critical for any similar real-world interpretation. We also cautioned due to syntax, polysemy, metonymy, and discourse in natural language, semantic similarity expressed by *NWD* is not an unambiguous parameter for meaning interpretations.

The present analysis of the Wirikuta case does not allow a better resolution of event development and its representation in the public sphere of the internet. A more reliable tool for precise hit counting, precise time-labeling and more frequent measurements over the time-range of the study should allow us to reach more detailed understanding of the dynamics of the internet and the usage of ICT by communities and their activist expressions.

The selection of classifier-terms defining a case of interest is crucial to the interpretation of its evolution over time: terms which turn out to show neutral evolution may be less indicative than terms with approaching or receding semantic similarity. Also, emphasizing cultural terms over economical, sociological or psychological terms as classifiers will lead to a corresponding orientation of conclusions and offer perspectives that may otherwise remain hidden.

Finally, while most of the key-terms and classifier-terms in our study have been chosen from the English language – as *lingua franca* of the internet and cultural globalization – a dedicated future study using Spanish classifiers to describe the Wirikuta case will give an insight into distinct semantic evolutions in a language-partitioned internet approach. Also a future repetition of our 1994-2013 measurements will verify whether the present findings are stable over time and as such assess the *NWD*-method and tree-ring growth-model of the internet. The effect on trend interpretation by selecting alternative classifier-terms and sample frequencies will then be studied as well. Furthermore, the temporal *NWD*-method should be tested in other cases of interest in order to compare its effectiveness and optimize its use.

The main result at this point is the observation that the *NWD*-evolution on the internet of key-terms and classifier-terms assigned to a case of social engagement does evolve with events over time, and thus can be applied as a meta-data based tool for measurement of online movement or activism. We conclude by summarizing the methodology of the temporal *NWD*-method:

(1) Determination of key-terms that unequivocally identify the case of interest by use of discourse analysis of available online information.

(2) Determination of positive and negative classifier-terms selected from polarizing discourses on the case of interest by use of available online information. These classifier-terms must be chosen to allow interpretation of the case-related research questions.

(3) Determination of one or more neutral classifier-terms which are invariably related over time to the key-terms, for negative control.

(4) Longitudinal measurement of *NWD* of all combinations of key-terms and classifier-terms in a series of constrained time windows according to the desired sample frequency.

(5) Trend analysis of positive, negative and neutral *NWD*s of key and classifier-terms.

Notes

1. Essentially the number N has to be a factor of a few powers larger than $n_y(u)$, $n_y(v)$ and $\mu(u \text{ AND } v)$: "[…] N which is the sum of the numbers of occurrences of search terms in each page, summed over all pages indexed […]" (Cilibrasi and Vitányi, 2007). But, it is stated "[…] This parameter N can be adjusted as appropriate, and one can often use the number of indexed pages for N." (Cilibrasi and Vitányi, 2007).
2. Only two measurements with very low joint counts n_{1994}(Huichol AND music) and n_{1994}(Huichol AND fashion) remain with NWD larger than 1 in the first year 1994 of our measurement – falling in the domain 1994-2000 of discarded data.
3. It is almost impossible to choose a term which does not "suffer" from too much polysemy; table can stand for "piece of furniture," "a format to display numbers" or "a flat surface," all of which are mostly considered as neutral. For example, mines can be used for "explosive devices" or "installations for mineral extraction" which in the present case tend to be perceived as negative or neutral.
4. Two missing joint query series were added in the period April 18, 2014-April 24, 2014, the numbers were in line with previous measurement of February and were added to the data set.
5. We noticed a nearly systematic estimation effect of hit counts by Google Search in date range modus for queries from 1994 till 2000 for terms with counts above approximately 100. For example the term peyote receives for seven consecutive annual measurements exactly the same number 16200 (measured on February 16, 2014, Brussels region).
6. For invariant time evolution – a horizontal line – the coefficient of determination R^2 by definition gives no qualitative indication.
7. For example, the PEN-international petition letter to president Felipe Calderon, following the *Grupo de los Cien Internacional*'s letter signed by 100 politicians, scholars, writers and artists (Grupo de los Cien International, 2011).
8. Federal court decision of 26/12/2012 concerning 38 concessions in Wirikuta. www.frenteendefensadewirikuta.org/?p=2528&lang=en
9. Some of these online documents became very popular, like the video "Wirikuta se defiende! Aho Colectivo" on YouTube, www.youtube.com/watch?v=YQcyxH9q55c. More recently the movie "Huicholes: The Last Peyote Guardians" (Vilchez, 2014) documents the present case. Artisanal product include traditional clothing, yarn thread paintings and beaded objects with mythical animal and peyote design.

References

AJAGI-Jalisco Association in Support of Indigenous Groups (2011), "Frente en Defensa de Wirikuta", available at: www.frenteendefensadewirikuta.org/ (accessed July 1, 2014).

Barnett, T.L. (2012a), "Canada meets Wirikuta: visit from Council of Canadians' Maude Barlow", *Huffingtonpost*, November 29, available at: www.huffingtonpost.com/tracy-l-barnett/canada-meets-wirikuta-vis_b_2154482.html (accessed July 1, 2014).

Barnett, T.L. (2012b), "Star power merges with spirituality at Wirikuta Fest", *Huffingtonpost*, May 30, available at: www.huffingtonpost.com/tracy-l-barnett/star-power-merges-with-sp_b_1549789.html (accessed July 1, 2014).

Castells, M. (2008), "The new public sphere: global civil society, communication networks, and global governance", *The Annals of the American Academy of Political and Social Science*, Vol. 616 No. 1, pp. 78-93.

Castells, M. (2012), *Networks of Outrage and Hope: Social Movements in the Internet Age*, Polity Press, Cambridge.

Castillo, C., Mendoza, M. and Poblete, B. (2013), "Predicting information credibility in time-sensitive social media", *Internet Research*, Vol. 23 No. 5, pp. 560-588.

Cultural Survival Org. (2013), "Good News – Mexico: mining concessions suspended in Wirikuta", available at: www.culturalsurvival.org/news/good-news-mexico-mining-concessions-suspended-wirikuta (accessed July 1, 2014).

Cilibrasi, R. and Vitányi, P.M.B. (2007), "The Google similarity distance", *IEEE Transaction on Knowledge and Data Engineering*, Vol. 19 No. 3, pp. 370-383.

Cilibrasi, R. and Vitányi, P.M.B. (2010), "Normalized web distance and word similarity", in Indurkhya, N. and Damerau, F.J. (Eds), *Handbook of Natural Language Processing*, CRC Press, Boca Raton, FL, pp. 293-314.

Enciso, L.A. (2012), "Ordenan a mineras suspender actividades en la zona sagrada Huichol de Wirikuta", *La Jornada*, February 28, available at: www.jornada.unam.mx/2012/02/28/sociedad/041n1soc (accessed July 1, 2014).

European Commission (2012), "Proposal for a regulation of the European parliament and of the council on the protection of individuals with regard to the processing of personal data and on the free movement of such data", COM 2012, 11 final, available at: http://ec.europa.eu/justice/data-protection/document/review2012/com_2012_11_en.pdf (accessed April 15, 2015).

Freyermuth, G.S. (2010), "Edges & nodes/cities & nets", in Sl Brandt, W.F. and Mehring, F. (Eds), *REAL – Yearbook of Research in English and American Literature*, Narr Francke Artempto Verlag, Tubingen, pp. 55-74.

Grupo de los Cien International (2011), "Writers and artists ask Mexico's president to cancel mining concessions in the sacred territory of the Huichol people", Letter December 1, available at: www.pen-international.org/wp-content/uploads/2011/12/CALDERON-letter-English.pdf (accessed July 1, 2014).

Huauxa Manaka Pact (2008), "President Felipe Calderón endorsed the five states pact for the preservation of Wixáritari culture", Durango, April, available at: http://calderon.presidencia.gob.mx/2008/04/el-presidente-calderon-en-la-firma-del-pacto-de-huauxa-manaka-para-la-preservacion-y-desarrollo-de-la-cultura-wixarika/ (accessed July 1, 2014).

Joseph, M.K. (2013), "Critical theory for women empowerment through ICT studies", *Qualitative Research Journal*, Vol. 13 No. 2, pp. 163 -177.

Jungherr, A. and Jürgens, P. (2013), "Forecasting the pulse. How deviations from regular patterns in online data can identify offline phenomena", *Internet Research*, Vol. 23 No. 5, pp. 589-607.

Kalampokis, E., Tambouris, E. and Tarabanis, K. (2013), "Understanding the predictive power of social media", *Internet Research*, Vol. 23 No. 5, pp. 544-559.

Kilgarriff, A. (2007), "Googleology is bad science", *Computational Linguistics*, Vol. 33 No. 1, pp. 147-151.

Lasica, J.D. (1998), "The net never forgets", *Salon*, November 25, available at: www.salon.com/1998/11/25/feature_253/ (accessed December 7, 2015).

Liffman, P.M. (2000), "Gourdvines, fires, and Wixárika territoriality", *Journal of the Southwest*, Vol. 42 No. 1, pp. 129-165, available at: www.jstor.org/stable/40170145 (accessed July 1, 2014).

Navarrete Linares, F. (2009), "Crisis and reinvention: the redefinition", in Singh, P. (Ed.), *Indigenous Identity and Activism*, Shipra Publications, Delhi, pp. 53-79.

Oliver, K.M., Wilkinson, G.L. and Benet, L.T. (1997), *Evaluating the Quality of Internet Information Sources: Consolidated Listing of Evaluation Criteria and Quality Indicators*, ED-MEDIA/ED-TELECOM, Calgary, available at: http://files.eric.ed.gov/fulltext/ED412927.pdf (accessed July 1, 2014).

Pariser, E. (2011), *The Filter Bubble: What the Internet is Hiding from You*, Penguin Press, New York, NY.

Pérez García, C.L. (2016), *Doctoral Thesis in Interdisciplinary Studies PhD (in preparation)*, Brussels Free University, Brussels.

Portal de Información Estadística y Geográfica para el Municipio de Catorce (2014), available at: www.catorceslp.gob.mx (accessed July 1, 2014).

Reyna-Jiménez, O.F. and Arce, A. (2014), "Heritage as a global counter-development strategy to fight transnational mining projects in Wirikuta, Mexico", oral report to XVIII ISA World Congress of Sociology, Yokohama, July 13-19.

Rheingold, H. (1993), *The Virtual Community: Homesteading on the Electronic Frontier*, Perseus Books, New York, NY.

Satoh, K. and Yamana, H. (2012), "Hit count reliability: how much can we trust hit counts?", *APWeb'12 Proceedings of the 14th Asia-Pacific International Conference on Web Technologies and Applications, ACM Digital Library*, pp. 751-758, available at: http://dl.acm.org/citation.cfm?id=2259221 (accessed December 7, 2015).

Schoen, H., Gayo-Avello, D., Metaxas, P.T., Mustafaraj, E., Strohmaier, M. and Gloor, P. (2013), "The power of prediction with social media", *Internet Research*, Vol. 23 No. 5, pp. 528-543.

United Nations Development Programme (2014), "Índice de Desarrollo Humano Municipal en México: Una nueva metodología", available at: www.undp.org.mx/IMG/pdf/IDH_municipal_PNUD.pdf (accessed July 1, 2014).

UNESCO (2004), "Huichol Route through the sacred sites to Huiricuta (Tatehuari Huajuye)", Tentative list", available at: http://whc.unesco.org/en/tentativelists/1959/ (accessed July 1, 2014).

UNESCO (2013), "Eighth session of the Intergovernmental Committee", Convention for the Safeguarding of the Intangible Cultural Heritage, Baku, available at: www.unesco.org/culture/ich/en/8com (accessed December 7, 2015).

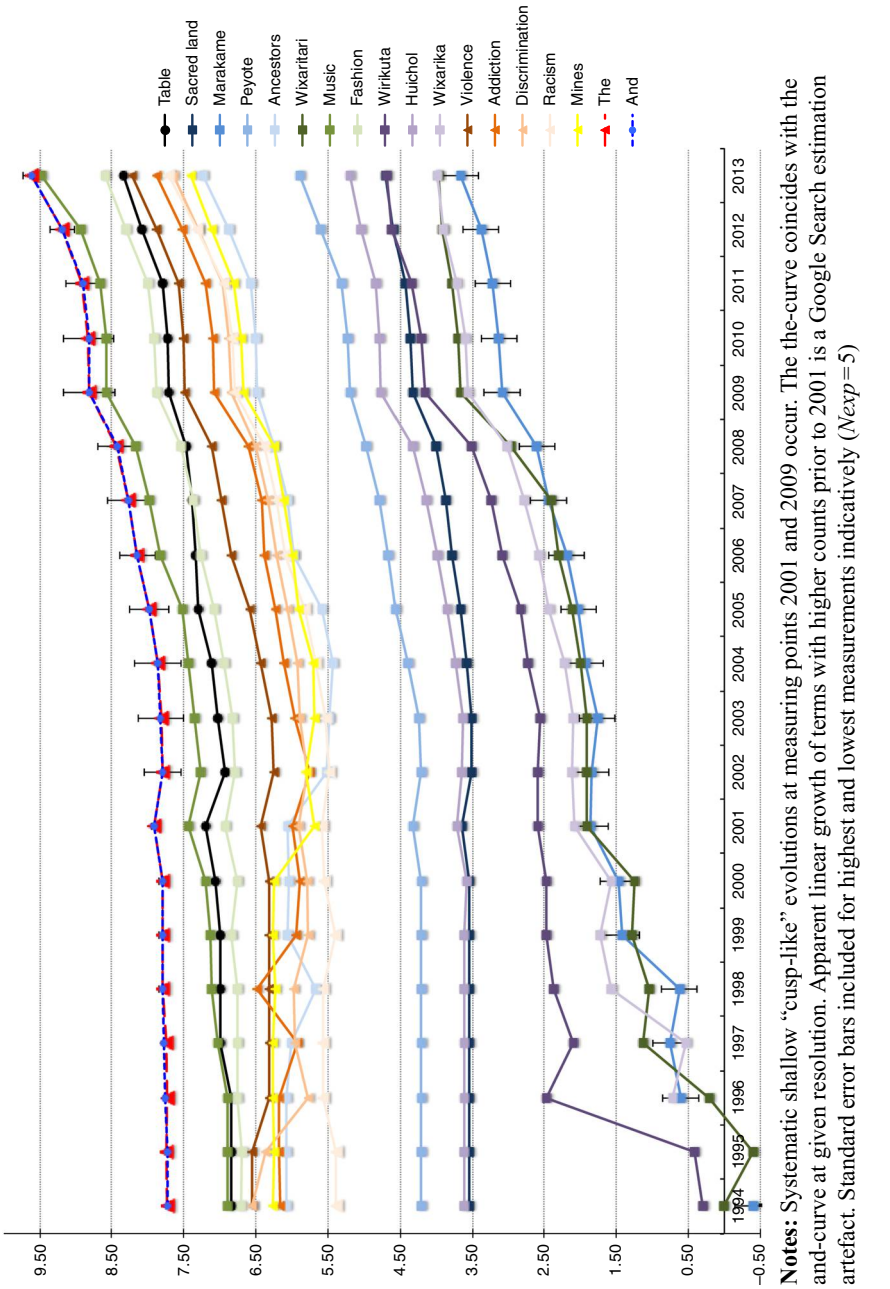
Vilchez, H. (2014), "Huicholes Los Últimos Guardianes del Peyote. (Huicholes: The Last Peyote Guardians)", Documentary, 120 minutes, available at: http://huicholesfilm.com/es/ (accessed July 1, 2014).

Webster, G.D. (2007), "What's in a name: is 'evolutionary psychology' eclipsing 'sociobiology' in the scientific literature", *Evolutionary Psychology*, Vol. 5 No. 4, pp. 683-695.

Further reading

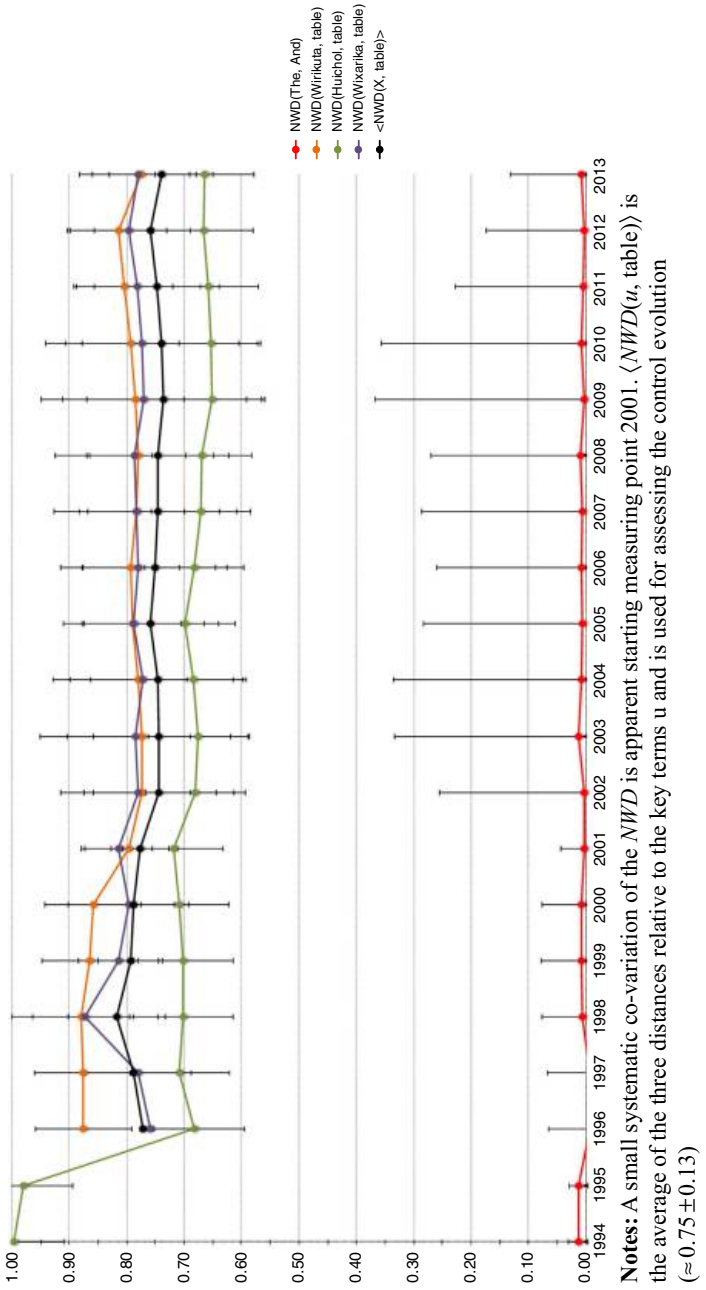
Liffman, P.M. (2011), *Huichol Territory and the Mexican Nation: Indigenous Ritual, Land Conflict, and Sovereignty Claims*, University of Arizona Press, Tucson.

Appendix 1. *NWD*-graphs



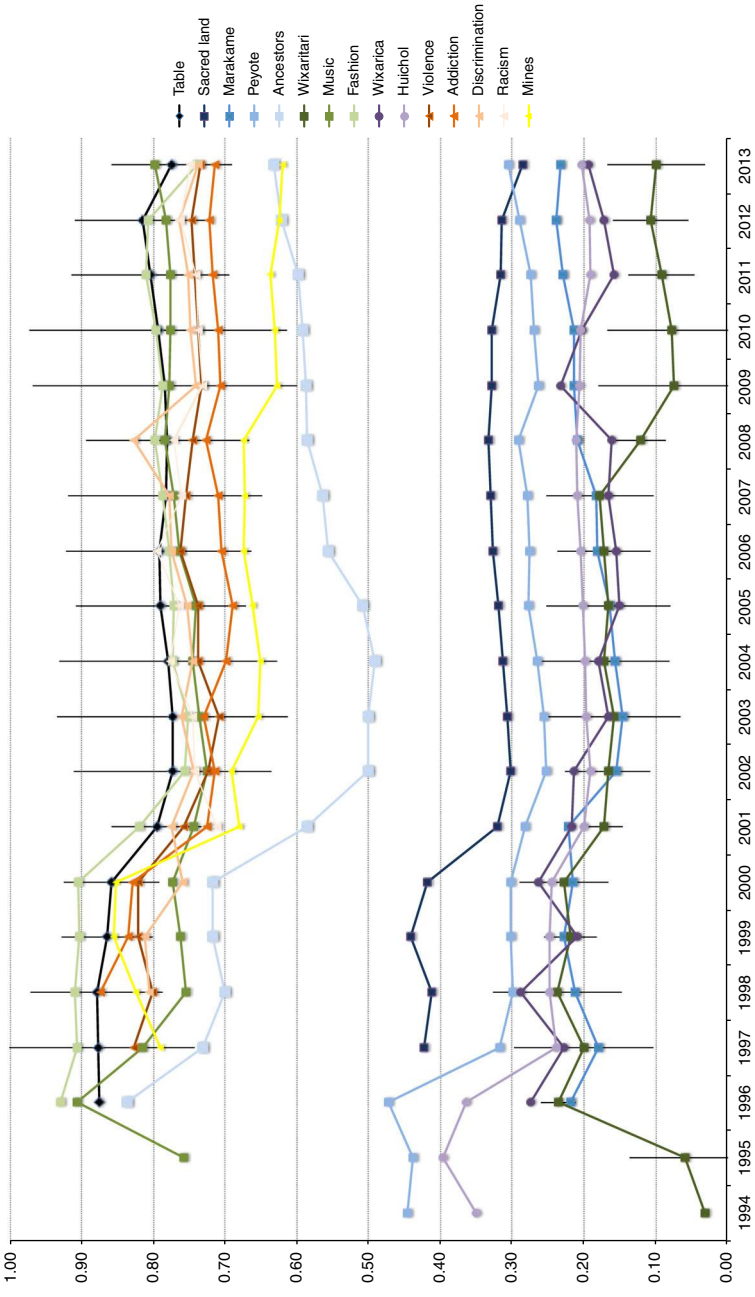
Notes: Systematic shallow "cusp-like" evolutions at measuring points 2001 and 2009 occur. The the-curve coincides with the and-curve at given resolution. Apparent linear growth of terms with higher counts prior to 2001 is a Google Search estimation artefact. Standard error bars included for highest and lowest measurements indicatively ($Nexp = 5$)

Figure A1.
Annual logarithmic incremental growth of page counts $\log_{10}(n_y(u))$ for term u



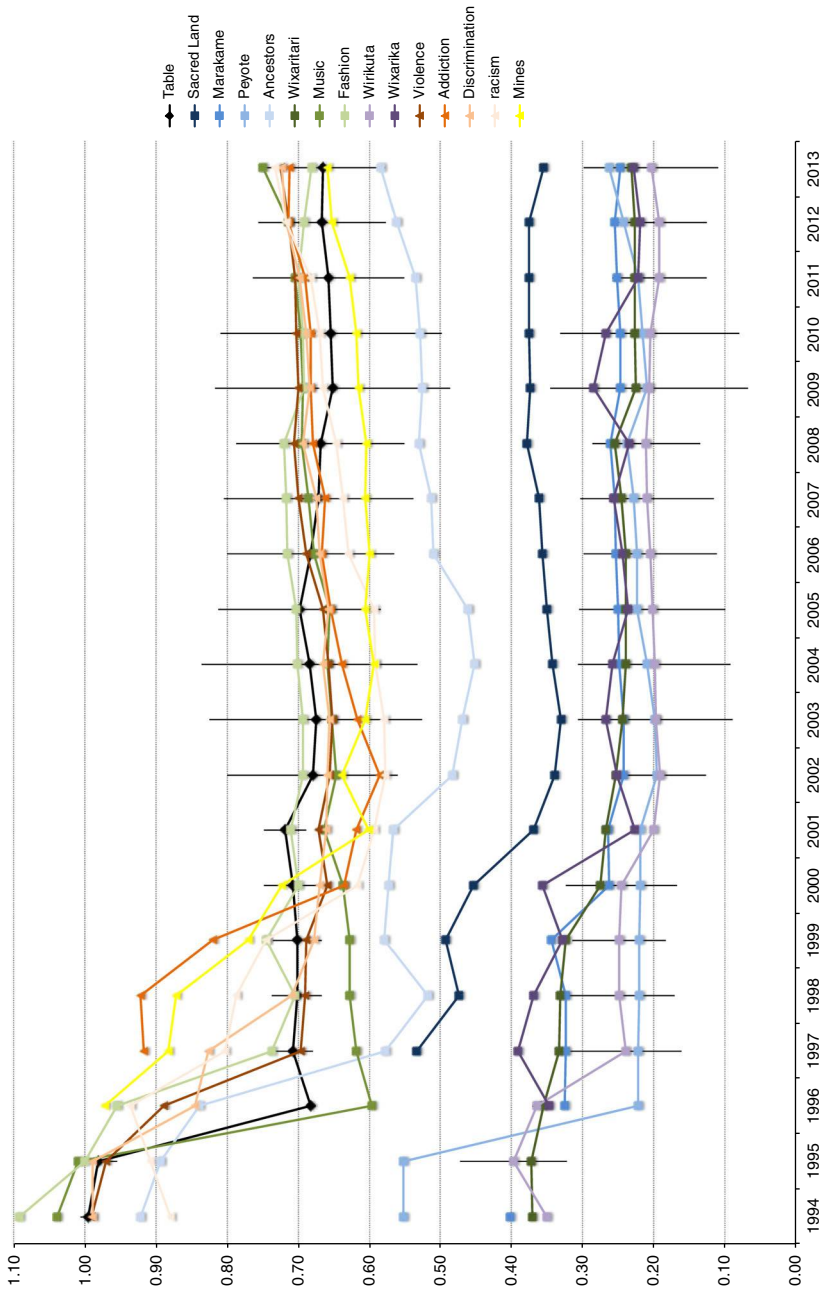
Notes: A small systematic co-variation of the *NWD* is apparent starting measuring point 2001. $\langle NWD(u, \text{table})\rangle$ is the average of the three distances relative to the key terms u and is used for assessing the control evolution ($\approx 0.75 \pm 0.13$)

Figure A2.
Annual change of control *NWD*'s; for *the* and *and* near zero, and for table with respect to Wirikuta (≈0.79 ±0.13), Huichol (≈0.67±0.12) and Wixarika (≈0.78 ±0.13)



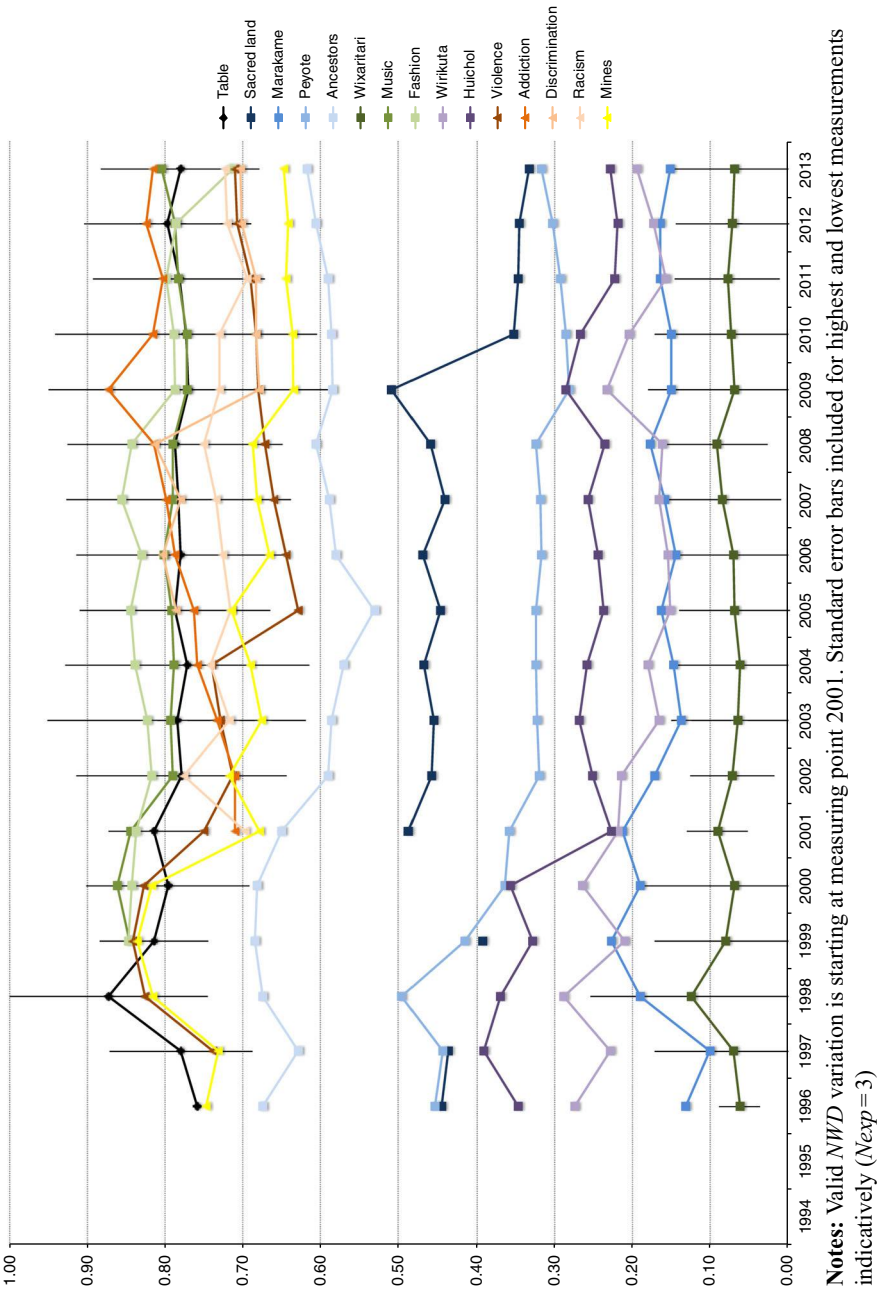
Notes: Valid *NWD* variation is starting at measuring point 2001. Standard error bars included for highest and lowest measurements indicatively (*Nexp* = 3)

Figure A3.
Temporal evolution of *NWD* (Wirkuta, *u*)



Notes: Valid *NWD* variation is starting at measuring point 2001. Standard error bars included for highest and lowest measurements indicatively (*Nexp* = 3)

Figure A4.
Temporal evolution of *NWD* (Huichol, *u*)



Notes: Valid *NWD* variation is starting at measuring point 2001. Standard error bars included for highest and lowest measurements indicatively (*Nexp* = 3)

Figure A5.
Temporal evolution of *NWD* (Wirkuta, *u*)

Appendix 2. Data on the Wirikuta-case with regard to internet sources

Geography

About 40,000 Huichol, or Wixaritari in their own denomination, are scattered in an area of about 5,000 km^2 in Western Sierra Madre in the central part of Mexico (Liffman, 2000). Of this region the major part is under indigenous control, while the remaining part stays culturally closely related to indigenous traditions. At the core of the Wixarika culture is their sacred land of pilgrimage, the Wirikuta. It is a cultural and natural reserve of about 140,000 hectares. The Wirikuta region extends south of the town of Real de Catorce in the state of San Luis Potosí, about 400 km to the East of their homeland. A detailed anthropological insight in the cultural relation of the Huichol and their pilgrimage to the Cerro Quemado in Wirikuta territory is given by Liffman (2000).

Events

Following prior protection measures – like the Huaxa Manaka Pact (2008) – by the local Huichol authority, the "Huichol Route through the sacred sites to Huiricuta" has been on the tentative list for UNESCO World Heritage since 2004 while the "Pilgrimage to Wirikuta" has been proposed as UNESCO Intangible Cultural Heritage by the Mexican government in 2013 (UNESCO, 2004, 2013; Reyna-Jiménez and Arce, 2014).

In 2010, the Mexican federal government granted open-sky mining concessions to boost economic development in the state of San Luis Potosí, a region known for small scale historic mining sites. This area is classified as highly marginal by the government, with a Human Development Index, that is among the lowest in the country (Portal de Información Estadística y Geográfica para el Municipio De Catorce, 2014; UNDP, 2014). In total, 500 direct jobs and 1,500 indirect jobs were envisaged along with a planned investment of over 17 million MXP or 1 million EUR approximately (*Huicholes: Los Últimos Guardianes del Peyote*, 2014).

The exploitation, however, would also take place within Wirikuta, one of the five sacred places for the Huichol community. By granting exploitation rights to the mining companies, the federal government would renounce the constitutional rights of land, property and self-government by the indigenous group. Mexico endorsed the rights of Tribal and Indigenous Peoples Convention 169 in 1992. In order to prevent the destruction of their cultural sites and natural deterioration due to mining, the Huichol and concerned supporters organized themselves to provide protection. They used social media to achieve their pursuit of international acknowledgment, mediating their voices, images, sounds and life experiences on the internet. During the second half of 2010 a resistance movement was started to halt the mining projects: the "Frente en Defensa de Wirikuta – Tamatsima Wahaa" (AJAGI-Jalisco Association in Support of Indigenous Groups, 2011). There were demonstrations against the government's actions, but the movement really came to prominence during the second half of 2011 with the support of NGOs. Information was recorded in Wixaritari and Spanish, translations in English, French, German and Italian followed. The spread of information over social media and the response of national and international organizations[7] and well-known activists (Barnett, 2012a) influenced the federal court's suspension in February 2012 of the La Luz mining project in Wirikuta until final arbitration (Enciso, 2012)[8]. Remaining concessions on the territory were however not affected. This situation spurred the online activism. Wirikuta Fests attracting crowds as large as 60,000 to the Foro Sol stadium in Mexico City were organized in 2012 and 2014 (Barnett, 2012b), short and long films along with fashion items were released to create awareness of the issue and to raise the funds needed to support the movement[9].

In September 2013, with the influence of the movement, the judicial power of San Luis Potosí at the level of the federal court suspended all remaining mining concessions – Universo and Maroma with 40 concessions – in Wirikuta (Cultural Survival Org., 2013). With this suspension, for the time being the sacred land of the Huicholes is protected and their traditions are sheltered.

Corresponding author
Jan Broekaert can be contacted at: Jan.Broekaert@vub.ac.be