Escola d'Enginyeria de Telecomunicació i Aeroespacial de Castelldefels
UNIVERSITAT POLITÈCNICA DE CATALUNYA

MASTER THESIS

TITLE : Cultural configuration of Wikipedia: measuring Autoreferentiality in different languages

MASTER DEGREE: Master of Science in Telecommunication Engineering & Management

AUTHOR: Marc Miquel Ribé

DIRECTOR: Horacio Rodríguez Hontoria

TUTOR: Sebastià Sallent Ribes

DATE: March 31, 2011

Títol : Cultural configuration of Wikipedia: measuring Autoreferentiality in different languages

Autor: Marc Miquel Ribé

Director: Horacio Rodríguez Hontoria

Tutor: Sebastià Sallent Ribes

Data: 31 de març de 2011

Resum

”Wikipedia és un projecte enciclopèdic multilingüe, col·laboratiu, basat en web i sense ànim de lucre impulsat per la Fundació Wikimedia”, així és com s’autodescriu Wikipedia en la definició de l’article que du el seu nom. Això significa que l’enciclopèdia pot ser modificada en qualsevol moment, per qualsevol persona i des de qualsevol lloc. Aquestes premisses i la seva gran participació fan que es tracti d’un excel·lent objecte social d’estudi, que a la vegada, per tractar-se d’un artefacte tecnològic, permeti també l’ús de tècniques de processament llenguatge natural, obtenció i mineria de dades. Tanmateix, en la recerca actual hi ha una clara mancança en software que pugui aproximar-s’hi d’una manera integral.

Tenint en compte aquest buit realitzem una caracterització de Wikipedia amb l’objectiu de conèixer a fons quins són els elements i estructures d’informació que conté i com després poden obtenir-se mitjançant una eina analítica. Partim de l’API existent anomenada wikAPIdia, que desenvolupem fins a incloure-hi noves funcionalitats i posar-la apunt per a encarar múltiples escenaris i problemàtiques de les ciències socials. Com a cas pràctic, revisem l’estat de l’art pel què fa a la motivació dels editors i la cobertura temàtica del repositori. Això ens permet considerar l’objectiu de conèixer Wikipedia des del punt de vista de tenir una configuració cultural única per a cada llengua. Formulat com a pregunta: ”existeix una motivació nacional o autorepresentativa que es reflecteixi en els continguts i els disposi diferenciadament?”

Autoreferencialitat és el concepte que presentem per tal d’analitzar aquest hipotètic interès superior reflectit en el contingut local. Realitzem una identificació i recol·lecció dels articles que poden referir-se tant a la història, als equips esportius o la cultura pop, entre d’altres, però que mantenen un lligam semàntic clar entorn l’àmbit dels editors. Posteriorment, en plantegem un anàlisi multidimensional fins a arribar a indicadors significatius, que permetin arribar a conclusions comunes i a la vegada compondre un índex per a la comparació. En darrer lloc, avaluem quin és l’impacte d’aquest contingut i la necessitat de que la seva existència sigui tinguda en compte en el disseny d’aplicacions basades en repositoris de coneixement que donin servei a usuaris.

Paraules clau: Viquipèdia, Desenvolupament del Software, Web Col·laborative, Cobertura Temàtica, Mineria de Dades.

Title : La configuració cultural de Wikipèdia: mesurant l'Autoreferencialitat en diferents versions lingüístiques

Author: Marc Miquel Ribé

Director: Horacio Rodríguez Hontoria

Tutor: Sebastià Sallent Ribes

Date: March 31, 2011

Abstract

"Wikipedia is a free web-based, collaborative, multilingual encyclopedia project supported by the non-profit Wikimedia Foundation" this is the way the definition of Wikipedia in the article of the English language edition starts. This means it can be modified at any time, by anyone and at any place. These bases and their participation success make of Wikipedia an excellent social object of study which, at the same time, for being a technological construct, can be approached by techniques of natural language processing, information retrieval or data mining. However, in the current research there is a clear lack of software which can make an integral approach.

Taking this into account, we make an in depth characterization of Wikipedia with the end goal of understanding which elements and structures compound its data and how they can be obtained with an analytical tool. We start with the existing API called wikAPIdia, which we develope until include new functionalities and have it ready to use in multiple scenarios and problematics of social sciences. Looking for a practical case to test it, we review the current state of art in motivation of editors and the topical coverage in the repository. This allows us to consider the aim of understanding Wikipedia from the perspective of having a different cultural configuration for each language. Phrasing it as a question: "is there a national or self-representative motivation which is reflected in the content and thus disposes them differenciately?"

Autoreferentiality is the concept we present in order to analyse this hypothetical higher interest in local content. An identification and recollection is made on articles from heterogenous topics which can refer to the local history, sport teams or pop culture, but still maintain a semantic relation to the context of editors. Later, we propose a multidimensional analysis of them on features which can be significant indicators, to reach common conclusions and evaluate the language editions through an index of Autoreferentiality. Last, we point out which is the impact of this content and the risk of not considering its existance in the design of applications based on user generated content.

Keywords: Wikipedia, Software Development, Collaborative Web, Topical Coverage, Data Mining.

To my beloved family. To Diana. To my friends.

”The goal is to give people a free encyclopedia to every person in the world, in their own language. Not just in a ’free beer’ kind of way, but also in the free speech kind of way.”

Jimmy Wales, Wikipedia co-founder

”But this emphasis would be lavished in vain, if it served, in your opinion, only to abstract a general type from phenomena whose particularity in our work would remain the essential thing for you, and whose original arrangement could be broken up only artificially.”

Jacques Lacan, Psychoanalist
Seminar on *The Purloined Letter*

CONTENTS

LIST OF FIGURES

LIST OF TABLES

PREFACE

Presentation

”Wikipedia is a free web-based, collaborative, multilingual encyclopedia project supported by the non-profit Wikimedia Foundation”, this is the way the definition of Wikipedia in the article of the English language edition starts. It means it can be modified at any time, by anyone and at any place.

The sixteen million articles made by all languages prove the participation success of a project which is considered to be the largest collaborative work in history. It started in 2001 and nowadays being ten years old it is one of the reference web pages on the Internet. Catalan Wikipedia takes the thirteenth position by number of articles among two hundred and seventy language editions. It was the third one to be created and the second one to have articles. The Catalan language Wikipedia, so called ”Viquipedia”, grew at a great speed, having a significant representation in number of articles compared to the number of speakers. Besides, it is also the first in quality articles indexed.

For its characteristics as a dynamic knowledge repository, Wikipedia has become an excellent object of study. Its nature is textual but at the same time relational, which makes it suit the requirements of academic fields such as natural language processing, data mining or interdisciplinary fields with data analysis.

The current technology developed to work with Wikipedia extensive XML data files, which includes all the information from the repository, is usually programmed in languages like Java, Python or Perl. It provides functions to extract the most important structures, but usually lacks any multilingual approach. Besides, there are no mechanisms to manipulate articles features and extract them for further statistical analysis. This is also the case of one of the most interesting research tools, wikAPIdia, from University of Northwestern, which struggles on the manipulation of sets of articles but it solves the multilingual question. Hence, considering this as an start-point, the goal is an improvement on this tool which will enhance the possibility of understanding in greater detail how Wikipedia presents changes across the languages.

Whether it is using the articles characteristics or its relations to the categories the repository can be approached from the quantitative perspective. Yet, it has been done, for instance, to understand the behaviour of editors which could turn into administrators or to detect vandalism. Studies which set a social context and model need at the end to extract data and verify their hypothesis.

Since the encyclopedia is available on editions from disparate social backgrounds and with different community sizes and activities, it would be no wonder that there might be some particular cultural configurations on the repositories. Either on the motivation which keeps the editors commited. In a non-profit project, understanding what involves a community and keeps its attention is a key question.

The answers from the academic about motivations types are fun, ideology, self-steem and reciprocity. Nevertheless, after reading a survey on users from the Catalan version [1] a new kind of motivation not treated by current literature emerged. There is a gap in the national or auto-representative motivation. Not only because of the inherent controversy in identity topics but also because every language edition treats in depth historical characters, cultural signs and social groups which have only a clear meaning to its speakers.

There might be a spontaneous interest on showing one’s own culture, language or activities. Encyclopaedias like German Brockhaus played an important role in defining a cultural point of view during the construction of the German national state, as well as the Catalan ’Enciclopèdia Catalana’ did, after the forty years of dictatorship, stating its goal ”promoure i difondre la llengua i cultura catalanes [...]”[2]” [4]”. In Wikipedia there is no ideological guideline for this to occur.

In this research, it is proposed to examine the articles related to this topic to understand how they are related to the language edition behaviour, its final influence and scope. For this it is required to develope a tool to approach multiple cultural questions and eventually define an analytical model for the empyrical analysis. This might be able to find commonalities across language editions as well as unique aspects.

It is necessary reviewing topical coverage in Wikipedia, whose main contribution is from Kittur et al. (2009),

[1] Primer, Segon i Tercer sondeig *Amical Viquipèdia*

[2] http://www.enciclopedia-catalana.cat/ca/fundacio.htm *Fundació Enciclopèdia Catalana*

these articles will be related to their topics but also to themselves as national, linguistic or autorepresentative of the language edition. A study from Hecht (2009) defined the concept Self-focus bias to analyse the prominence of geographical articles related to the language edition.

Here an alternative understanding of the concept's definition is proposed as Autoreferentiality. In this operalization there will be indicators on what may affect or influence its reason of being, the interest on these articles as an effect of a new kind of motivation. Finally, after checking its hypothesis, there will established a groundtruth for a comparative index across all the tested language editions.

The comparative study will be extended to a set of twenty languages and have a dedicated section for the Catalan version, which will also receive more attention during the theoretical part. The languages will be diverse, selection criteria will comprise: territory expansion, linguistic roots and region, number of speakers and editors. This will increase its validity and help avoid the common mistake in Wikipedia research of extrapolating results from English edition.

Thus, firstly there will be a source data extraction in order to import the main structures into a database and have them ready for prior analysis and operations. This will imply dealing with bottlenecks in memory and disk spaces due to the size of the files to be parsed. The programming framework and work scenario will require to be adapted to this purpose and find scalable solutions to the new extensions.

Also, a new abstraction on what is an article, a set of articles, its selection processes and criterias are key questions to be treated. Afterwards, once analysed their features and checked the hypothesis, there will be mechanisms to extract selected articles or characteristics which will enable the possibility of continuing the examination by statistical or data mining means. Albeit these last will require a proper syntaxis in the files to be considered by the tool output methods. Last, the current study will also be presented on a wiki webpage. For collaborative reasons it is the best way to spread the study and its results among the Wikipedia community and a call for its continuous improvement.

Structure

In this document, the contents are distributed in four chapters: *Introduction*, *Approach*, *Design and implementation* and *Conclusions*, and two appendixes on Results.

The ***first chapter*** introduces Wikipedia in its technological and social perspectives, its working functioning and the multilingual or hyperlingual nature. Thereafter it will present the state of art, regarding the analytical tools and purposes of Wikipedia, as well as cultural influence on motivation and topical coverage.

In the ***second chapter*** called Approach one may find the proposal of research. A prior part will entable how a technical characterization describes the Wikipedia structures. Moreover, Autoreferentiality is established as a multidimensional case on which to apply the tool. How to obtain the autorepresentative articles for the study, its criterias, however, will also be detailed.

As a ***third chapter***, Design and Implementation, will detail how the technical characterization and analysis have been developed into actual modification in the API. For this it will set requirements which will be applied in particular problems. It will be necessary to dig into the Framework API and database to present its extensions. Later, there is a section based on testing the cost or performance of the tool in multiple scale files and processes.

The ***fourth chapter***, called Results, will evaluate the indicators and their related hypothesis with the end goal of presenting an index. There will be a discussion on each analysed indicator to understand their influence and later they will be tested using correlation to see if there is redundant information. The index will show the degree of Autoreferentiality of the language editions.

Fifth and last chapter, the Conclusions, will be divided in two different parts. Discussion will develope some of the ideas around the obtained results, achieved goals will be basically an evaluation on how the development have been carried out, conclusions will wrap up both points of view of the research and finally in future future goals there will be some interesting future lines.

The ***Appendix*** will include a chapter called, Complementary Results, which will develope some other indicators and in depth analyses.

Impact and contribution

Some of the results and knowledge obtained in the process of this research have been presented in several events and publications related to the Wikipedia community and the academic research.

- TFC "La configuració cultural de Wikipedia: estudi de l'Autoreferencialitat en diferents versions lingüístiques a Wikipedia" Universitat Oberta de Catalunya - Invited presentation to the cerimony "Humanistes del 2011". February 17th, 2011. Barcelona, Catalonia.

  A preliminar version of this study was presented in the context of the graduation ceremony with two more selected works among 111. It included eight different Wikipedia language editions and less features than this work. The theoretical part, scope and key words methodology were validated. The work obtained Honors mark.

- "Cultural Influence on Knowledge Repositories for Semantic Web" in COST 2011 Conference on "Cross Modal Analysis of Verbal and Nonverbal Communication" held in Dresden Technische Universität. 21-25th February, 2011. Dresden, Germany. Published in Lecture Notes in Computer Science (LNCS).[3]

  In the context of a multidisciplinar conference on verbal and non-verbal communication some results of this work were presented to remark the cultural differences on the web. The COST action which organized the event consider the web semantic as a strategic point.

- Participation on the seminar about "Collaborative Web" by MediAccions IN3 PHD Program. March 1st, 2011.[4]

  IN3 Interdisciplinary research group Mediaccions organizes periodical seminars in which this work was introduced and discussed next to thesis on Wikipedia governance and other web/digital media production studies. The developed tool raised some interest on future collaborations.

- Amical Viquipèdia - Round table around Wikipedia "Viquipèdia i Recerca". March 15th, 2011. Barcelona[5]

  Catalan Wikipedia friends association celebrates the Viquipèdia tenth anniversary and organizes an event to discuss the current research on Wikipedia. This work is presented next to other contributions.

- Wikimeeting International (promoted by Wikimedia CAT) - "Motivations and Interests on Wikipedia: understanding Viquipèdia to promote it" - March 17-20th, 2011. Perpignan, North Catalonia.[6]

  Wikimedia CAT organizes an international event in a development and promotion Plan to open new contacts with other Wikipedia languages and find new means to expand the communities. This work analyzes "euskera", "ex-URSS languages", "occitan" and "aragonese" in order to understand their current situation.

- Wikimania 2011. "Ethnocentrism in 25 Wikipedia languages". August 5th, 2011. Haifa, Israel.[7]

  Part of this research will be presented in the annual world-wide Wikipedia conference Wikimania supported by the Catalan Amical Wikipedia association. Its goal is to explain the differences in repositories from the analysis. There will be an extra emphasis on hebrew language since the event is hold in Haifa.

This study will be presented separately in different academic contributions showing two problematics: the scope of local semantic content on Wikipedia languages and its multidimensional analysis.

- "What's local in Wikipedia? Study on language related content". On RANLP 2011.

[3] European Cooperation in the field of Scientific and Technical Research - http://www.ias.et.tu-dresden.de/ias/fileadmin/user_upload/sprache/COST2102/Schedule-v6.pdf

[4] Semanari Mediaccions - http://mediacciones.es/seminari-mediacciones-cinema-digital-y-wikipedia/

[5] Round table "Viquipèdia i Recerca" program - http://ca.wikipedia.org/wiki/Viquip\%C3\%A8dia:10_anys_de_la_Viquip\%C3\%A8dia#Taula_Rodona:_Viquip.C3.A8dia_i_recerca

[6] Wikimedia CAT celebration and future lines event - http://ca.wikipedia.org/wiki/Viquip\%C3\%A8dia:10_anys_de_la_Viquip\%C3\%A8dia#Viquitrobada_internacional

[7] Wikimania Submissions - http://wikimania2011.wikimedia.org/wiki/Category:Wikimania_submissions

GOALS

To summarize how this study will approach the Wikipedia object and solve a social question it is useful to set general goals which later in turn will be defined in characterizations and solutions. Firstly, the scope requires a **primary goal** (technical) as *projecting, developing and testing a tool which enhances the possibility of extracting information from Wikipedia and know better its complexity*.

Thus it needs to take into account the six following requisites:

- Analysing Wikipedia as a knowledge repository to understand how it can be conceptualized.
- Abstracting the Wikipedia repository in a programming framework as a relational construction.
- Obtaining and storing characteristics from Wikipedia structures in a scalable information system.
- Operating with sets of data, being them articles, categories or properties.
- Solving the multilingual nature of WP from both compatibility and resources consume.
- Providing mechanisms to select and extract data to proceed with statistical and mining assessment.
- Exporting results into different formats which enable data representation and its interpretation.

Thereafter, when analysing Wikipedia the **secondary goal** (scientific) is to understand the *cultural configuration of Wikipedia repositories or how a possible national or autorepresentative motivation can manifest differently in their language editions articles*. For this, in any language edition there will be a selection of articles from heterogeneous topics around the national or linguistic sphere.

The comparison of this set of articles characteristics in relation to those which present all the articles from a language edition of Wikipedia should allow seeing indicators of a higher interest in the first. *Autoreferentiality* is the concept which comprises all the different dimensions of analysis and is after a measurable property of a language edition.

Those confirmed indicators showing interest will be revised to see its value as information and avoiding redundancy. At last, an index to be applied to all the set of languages will be developed. Therefore, in this procedure it is essential to identify and discuss factors like:

- The interest aroused by articles in other language editions.
- The dedication or effort from editors on articles' content.
- The importance or prominence of articles concerning its relations with others.
- The historical dimension of the articles as a sum of edits by editors.
- The temporal trends on article creation and its behaviours.

It is considered appropiate regarding the catalan edition Viquipèdia, to set the two **specific goals**:

- Ellaborate a ranking of articles from the selected set.
- Describe a topology of articles prioritizing characteristics.

Last, taking into account the social dimension of the study a **third goal** (social) will be to prepare the tool and environment to *extract the information in a friendly format for the same Wikipedia communities*. It seems clear that the more awareness the editors have about their creation the better they can continue their work.

CHAPTER 1. INTRODUCTION

”Given enough eyeballs, all bugs are shallow”, Eric S. Raymond.

1.1. Wikipedia

This first chapter presents Wikipedia from both social and technological perspective as well as the current state of art in research. First of all there is an introduction to the repository regarding its social magnitude, including policies and multilingual characteristic. Then a description on how the encyclopedia is implemented in the Internet as a wiki-based webpage, its information and data formats.

Taking this into account, a technical research approach to Wikipedia might relate to three methodologies such as Information Retrieval, Data Mining and Natural Language Processing. The current tools are described and reviewed in order to create or extend one to our purposes as goal of this project. Later, from the current studies on motivation and topical coverage the social problematic context is defined.

1.1.1. Free, web-based, collaborative, multilingual encyclopedia

In order to introduce Wikipedia the most useful presentation is given by its name-article in the English version, which describes it as a ”free, web-based, collaborative, multilingual encyclopedia project supported by the non-profit Wikimedia Foundation”.[1]

The encyclopedic project was started on January fifteenth in 2001 as bifurcation of a free culture expert encyclopedia called Nupedia. Once they decide to open it to an altruistic collaboration based on anonymity, the story of it changes completely. What in free software was the freedom in obtaining, using, modifying and delivering the code, in Wikimedia Foundation texts meant modify, reproduce and copy regarding the copyleft license conditions.

Wikipedia started as a collective creation webpage before the user being the principal information producer as it is nowadays. Nevertheless, the difference between Wikipedia and another user-content-generated repository is its ideological background, goals and policies. Its funding is completely transparent and it is generated by individual contributions. In the 2008-2009 report they counted 8,6M dollars.

It amazes how a product designed in this way can grow under these conditions and be the seventh most visited webpage on the Internet according to the Alexa ranking[2] [1]. It is only preceeded by social networks, portals or web searchers developed by big companies. The collective achievement is produced by 25 million registered users and has over 16 million articles, in 275 languages out of which 35 have over one hundred thousand articles. These numbers make Wikipedia the biggest database created in a collective work.

Regarding the format in which the content is presented, Wikipedia implements the concept *wiki* [24]. It was designed in 1994 by the American programmer Ward Cunningham. Its name comes from Hawaiian wikiwiki (fast) and in webpages applies to the ease in creating or modifying text collaborively as well as the property of linking words to other articles.

It was implemented from the start despite some reluctance by co-founder Jimmy Wales. Later, other features were added such as the history of editing or discussion areas that are essential for the use of Wikipedia.

1.1.2. Governance

The ideological purpose of spreading free knowledge, wiki technology and open collaboration of anonymous subscribers are the basis for the functioning of Wikipedia. To achieve reliability and good results rules

[1] http://en.wikipedia.org/wiki/Wikipedia - article ”Wikipedia” in english edition

[2] http://www.alexa.com/topsites *Alexa Top 500 Global Sites*

of behavior are drafted and developed in different documents: *Foudation Issues* in Wikimedia Foundation projects, *Statement of Principles* [22] from Jimmy Wales, and *Five Pillars of Wikipedia*. The following four fundamental aspects are common in them:

- Neutral Point of View. Is the top editorial position in regard to the writing. An article should not be a point of view but a search for objectivity, and must provide a representation of various perspectives in each context. To achieve this, it also requires precision references. When a dispute or conflict occurs and it endangers the text, the discussion pages are used and it ends with a resolution of the dispute.
- Free content that anyone can edit. It means that the articles have no authorship and that any page can be modified at any time by anyone - even anonymously. It has to be clarified that, in spite of the spontaneous emergence of vandalism, Wales said that he does not want to implement a strict security and identifying system and in the community new users are always welcomed.
- Lack of fixed rules. There are no fixed norms besides some technological conditions (such as the impossibility of one text to be edited by several people at the same time). The same consensus is on establishing a dialogue on what is most convenient for the encyclopedia, and literally, if a rule prevents one from improving or maintaining Wikipedia, it is recommended to be ignored. In the edition it usually produces the acceptance of new elements introduced and otherwise, it reverts.
- A code of ethics. The interaction must be civilized and respectful, in exceptional cases, where vandalism is clearly observed in the material and the discussion, a publisher can be banned. The diplomacy Wikipedia wants for its editors consists of moral principles such as honesty and politeness.

In the English version, only in metapages, informative articles about what is recommended and not recommended, there are 344 articles that are linked to categories derived from *Wikipedia Guidelines* - regardless of user templates - and 56 to the category *Wikipedia policy*. These rules are continually updated and new proposals undergo reviews and are even erased. Butler et al. (2008) [3] suggested in a study that these are essential for the daily activities of the community. In particular, there are rules for behaviour, content, deleting, policies and procedures. It must be taken into consideration that, using consensus makes it highly possible for rules to appear particular to each version. In its application it keeps the fixed structure of users and their permissions for all versions.

The hierarchy of permissions sets on the lowest level the *anonymous*, user, who can view and edit all the pages, less those protected or semi-protected. *Registered* users have access to the features of the community and the *confirmed* users have more editing permissions. The profile functions completely changes for *administrators* who can delete pages, protect them or block them, among others. More focused on editing and activation of users is the *bureaucrat*. While on the last level, the *Steward* is a post elected out of all projects around the foundation, he may revoke any permit for user access to all past practices. Jimmy Wales, who had that status, recently has been recognized as a founder, or, as the editors refer to him, *Godking*.

1.1.3. Hyperlingua and Viquipèdia

Although the precursor Nupedia was intended to be only in English, Wikipedia had the debate of bifurcation from the beginning. Once passing the two hundred articles threshold, on March 16th 2001, Jimmy Wales announced the creation of other language editions. The story goes that the e-mail spreading the proposal, also mentions the special interest of a user, *Cdani* in the existence of a Wikipedia in Catalan ”for not to infringe his terrible English on them.”

On the same day and only minutes apart two new domains, *deutsche.wikipedia.com* and *catalan.wikipedia.com* were created. The first non-English language was made the front page of the Catalan version, at 21:07 UTC, and the first contribution, the next day at 1:41 UTC with the article Abacus. A year later, it already had sixteen editions and the number has not stopped growing ever since.

The Catalan edition of the Wikipedia changed the domain *ca.wikipedia.com* and afterwards to *ca.wikipedia.org*. The adoption of the name *Viquipèdia* was not until two years later, through an agreement between users active at the time. Different reasons were put forward, such as the influence of Vikipedio - name of the

Esperanto version - or the normal adaptation of w from English to Catalan. Other language versions also have the word adaptation, recognized by the foundation.

Another remarkable debate, in 2005, was on the use of the name of the language. A consensus was not reached at the end of all proposals, but it ended up accepting that the ”Viquipèdia in Catalan” could also be ”Viquipèdia in Valencian”, in its proper context. The book-style version allows any of the dialects recognized by the criteria of the IEC or AVL[3], although it requires that an article is in the same range and those relating to geographic or cultural areas are made with the local dialect.

The highlight of the Catalan Wikipedia is clearly reflected in the overall figures [23]. Significant is ranking 13 in number of articles. On December 21st 2010, it exceeded 300,000 articles, and today it counts 303,207 (the English version being first has 3,529,781). The number of administrators is low (27) when compared with the following version, the Norwegian, which has eleven thousand articles more and 69 administrators, and the Chinese version, which has forty-one thousand articles more and 76 administrators.

Regarding the number of users, it gives the same type of data: 77,609 for the Catalan version, compared to 184,845 for the Norwegian and 944,403 for the Chinese. As a reference, the English in this case has 13,767,765. Data users are reduced only when counting those assets of at least five editions, it drops to 1535 for the version Catalan, to 5137 for the Norwegian and to 2097 for the Chinese. If you go up the scale to a hundred editions, the Catalan version has 98 publishers, the Chinese 282 and the Norwegian 75.

In short, when wanting to know the activity of a linguistic community, one of the important data is the number of publishers for millions of speakers. The Chinese has 1, the Norwegian has 117, the English 24 and the Catalan 44. This difference is not scalable, but explicative about the community of Viquipèdia, also has other attributes. If we look at the length of articles, those exceeding 2 Kb are 35% for the version in Catalan, 23% for the Norwegian and 16% for the Chinese.

The Catalan Wikipedia counts with a high activity level. Also to be emphasized is the first position in the rankings for the quality of the 1000 most important articles of the encyclopaedia - it currently retains a score of 90.17 out of 100, above the English (86.48). Maintaining a high level of accuracy of the fundamental articles is a major objective for the future.

1.1.4. Technology

The current implementation of the wiki is called MediaWiki. It was launched in Perl and later it changed into PHP - being always based on a MySQL database. It was adopted as the technology for all Wikimedia Foundation projects and it is nowadays very popular among all the wiki implementations. It is also free software under the GNU General Public License (GNU GPL). For its goal it includes a list of automations which enhance the work on editing and managing.

On Figure 1.2 one can see a random Wikipedia article screenshot and on Figure 1.1 all the Wikipedia structures. Users navigate across the content using its links, category memberships and its searcher, like a typical web. Nevertheless, there are some unique elements which may need further attention.

For instance, any article has text which can be divided into snippets (first paragraph defines the concept). The article text or Wikipedia Text (aka WT) uses the hypertext quality of linking words into other pages from the same Wikipedia. Then it implies the repository must be complete to define itself. When links are related to a word from an article text to another article, they are called outlinks. From the cited article are called inlinks.

[3]language normative institutions Institut d’Estudis Catalans and Acadèmica Valenciana de la Llengua *www.iec.cat - avl.gva.es*

	Data Source	Description
1	**Article page text**	also referred to as Wikipedia Text (WT)
1a	Title Hierarchy	short definition of context, identification of pure temporal articles
1b	Snippets	Paragraphs within the full-text of the article
1c	First Snippet	gloss, or short definition, of subject of Wikipedia article
1d	Temporal references	pointers to the temporal reference system
2	**Redirects**	information on synonyms, spelling variations, common misspellings, common case variations
3	**Other page spaces**	
3a	Images	picture to represent the context
3b	External links and bibliography	list and link resources
3c	Textbox	sorts context information as characteristics
3d	Similar pages	suggests context related pages
3e	Spatial references	point references to the WGS 1984 geographic coordinate system
3f	History	lists all the revision and changes since its creation
3g	Discussion	comment based space about the page content
4	**Wikipedia Article Graph (WAG)**	
4a	Links	backbone of graph between articles
4b	Link label	information on synonyms, spelling variations, related terms
5	**Universal WAG (UWAG)**	
5a	Interwiki Links	backbone of graph between wikis and articles
6	**Wikipedia Category Graph (WCG)**	
6a	Category	contains category links, sometimes short descriptions of subject of category
6b	Category links	backbone of "folksonomy" between categories, contains almost entirely "isA" and "hasA" relations
6c	Category memberships	locates articles within a category "folksonomy"
7	**Disambiguation pages**	
7a	Disambiguation links	sense inventory
8	**Templates**	
8a	Wikiprojects	context-appropriate structured information
8b	Portals	context-appropriate structured information

Figure 1.1: An expanded version of a table in Zesch et. al (2007) which describes Wikipedia data sources

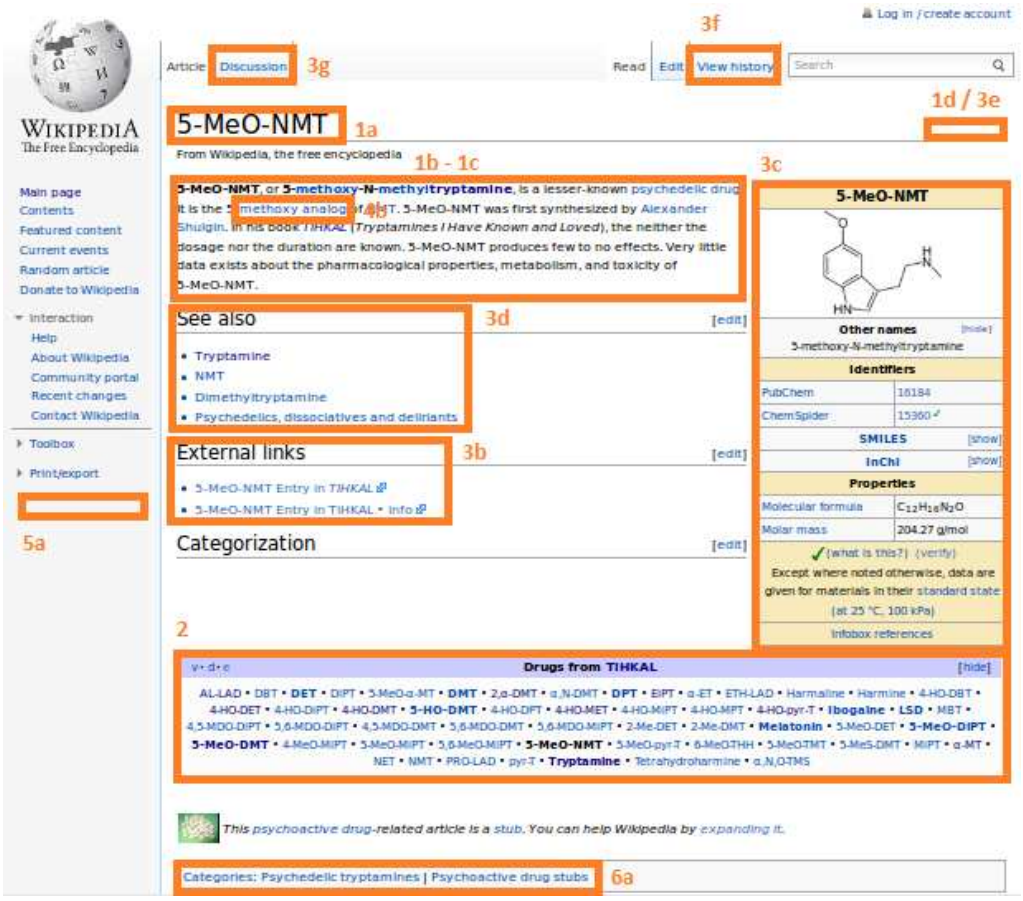

Figure 1.2: A random Wikipedia article screenshot

When articles exist in different language editions they are linked with interwiki links (or ILLs, interlanguage links). Also, an article may include several forms of information to enrich the content: temporal references, images, textboxes with characteristics or spatial references. All of them are subject to changes across time and can be recovered in the history page which keeps all the edits.

Not necessarily, but articles can be created out of a Wikiproject (where several editors have structured a field or theme) and can be included in Portals, which have the same contextual function but towards readers. Neverthless, the main classifying structure is a category. Articles and categories are usually members of categories, which define the topic or context in a broader sense.

Links between articles and links between categories can respectively create a Wikipedia Article Graph (WAG) and a Wikipedia Category Graph (WCG) where those entities can be understood as nodes. Taking into account the interwiki links, the coincident articles among different languages can can create a Universal Wikipedia Article Graph (UWAG).

From the server side Wikipedia started as a centralized cluster located in Tampa, Florida (USA), and moved to San Francisco, where there are the current Wikimedia Foundation headquarters. After, a Nederlands ICT company started providing service for all european countries and Yahoo donated a cluster in Seul (South Korea). That aproximately makes a system of 30 nodes for MySQL databases, 60 nodes for Squid proxy and web caching servers, 180 for executing Apache web servers and 20 additional for miscellanous purposes.

These last type of nodes provide also data collections or data dumps, which makes it unnecessary to use crawling techniques to gather data. Besides it is a practice banned, all backup dumps are provided regularly to the public every two weeks. They come as XML files, keeping its original syntaxis but without images, and for all Wikimedia Foundation projects and languages.

The history file (which includes all the edits respect the ”last articles” file) tend to be sized as several dozens of GB for those languages over hundred thousand articles. The ”last articles” is the most used. Both come compressed in various formats. Wikimedia Foundation provides these data collections for several purposes like archival, backup, offline use, republishing or academic research. On Figure 1.3 there is a piece of the article ”Anarchism” from the last articles file. In it there are the main data structures which identify an article and are previously explained in this section: title, id from the article, timestamp for the edit, contributor username and id, redirection link, and finally the text.

```
<title>Anarchism</title>
  <id>12</id>
  <revision>
    <id>149030244</id>
    <timestamp>2007-08-03T23:24:05Z</timestamp>
    <contributor>
      <username>Jacob Haller</username>
      <id>164072</id>
    </contributor>
    <comment>/* Four monopolies */</comment>
    <text xml:space="preserve">{{dablink|"Anarchist" redirects here. For the
    comic book character, see [[Anarchist (comics)]].}} {{toolong}} {{disputed}}
    {{Anarchism}} '''Anarchism''' is a [[political philosophy]] or group of
    philosophies
```

Figure 1.3: Part of the article ”anarchism” in the XML tag format

1.2. Related research

Some technical approaches have been perpetrated to solve some of the previous questions. The use of processing techniques to obtain valuable metainformation have been used with proved success. The presented approach will propose and develope an existent tool to know the reality of Wikipedia from its complexity. This is wikAPIdia, initially developed by researchers from the University of Northwestern.

In order to test it in a problematic which satisfies the interest of the social research community the motivation of users and topical distribution coverage issues of Wikipedia will be reviewed. Hence, understanding those two factors, the scenario for a study on the cultural influence or configuration on different Wikipedia language editions is set.

1.2.1. Technical approach

Related academic research is generally vast in many areas since Wikipedia is a knowledge repository which has been developed for more than ten years. Interdisciplinary studies have found in Wikipedia a resource of real information in multiple languages which can be an excellent object to solve human-computer interaction questions and study the creation and evolution of text on different topics.

These could lead to the use of techniques related to the semantic relatedness between articles Hecht [11] or the analysis of Wikipedia categories structure like a folksonomy. Nevertheless, summarizing the areas which refer to the use of its data for the processing and extraction of features in technical methodologies one find mainly three specific fields. These are Information Retrieval, Data mining and Natural language processing (NLP), also briefly explained in their Wikipedia articles [24].

- Information retrieval (IR) is the science of searching for documents, for information within documents, and for metadata about documents, as well as that of searching relational databases and the World Wide Web. There is overlap in the usage of the terms data retrieval, document retrieval, information retrieval, and text retrieval, but each also has its own body of literature, theory, praxis, and technologies. IR is interdisciplinary, based on computer science, mathematics, library science, information science, information architecture, cognitive psychology, linguistics, and statistics.

  Automated information retrieval systems are used to reduce what has been called "information overload". Many universities and public libraries use IR systems to provide access to books, journals and other documents. Web search engines are the most visible IR applications.

- Data Mining (DM), a branch of computer science, is the process of extracting patterns from large data sets by combining methods from statistics and artificial intelligence with database management. Data mining is seen as an increasingly important tool by modern business to transform data into business intelligence giving an informational advantage. It is currently used in a wide range of profiling practices, such as marketing, surveillance, fraud detection, and scientific discovery.

  The related terms data dredging, data fishing and data snooping refer to the use of data mining techniques to sample portions of the larger population data set that are (or may be) too small for reliable statistical inferences to be made about the validity of any patterns discovered. These techniques can, however, be used in the creation of new hypotheses to test against the larger data populations.

- Natural Language Processing (NLP) is a field of computer science and linguistics concerned with the interactions between computers and human (natural) languages. In theory, natural-language processing is a very attractive method of human-computer interaction. Natural language understanding is sometimes referred to as an AI-complete problem, because natural-language recognition seems to require extensive knowledge about the outside world and the ability to manipulate it. NLP has significant overlap with the field of computational linguistics, and is often considered a sub-field of artificial intelligence.

Normally in any of these cases WT (Wikipedia Text) is parsed from articles and studied. Studies which analyze Wikipedia like an interdisciplinary context to build their problematics and then obtain other information referring the structures, metadata of the articles which can be edition history or discussion. Papers on conflict detecting or administrators recruiting are found in this category.

Neverthless, the tools which are deployed until the moment present some necessities when looking at the table of Wikipedia data sources from last section. Those interacting with the XML dumps are usually programmed in languages like perl, python or java. Specially the two first languages, are used in text processing, but also as viewers for other Wikipedia-based applications.

Java-based software, MWDumper, is a quick little tool for extracting sets of pages from a file and output back to XML or to SQL statemenets to add things directly to a MediaWiki database. Also, MWImport in perl can parse faster than the previous one but may not work in all cases. Other options may be Xml2SQL or pywikipediabot in python but they are not able to parse certain MediaWiki tags like "#REDIRECT".

WikiXRay [15] is project developed in a phD thesis "Wikipedia: a quantitative analysis". Its only goal is an in-depth quantitative analysis of the Wikipedia project. It provides a python tool to process dumps for research

purposes capable of loading pages, edits and text tables from any language edition and some meta-data regarding the number of words, letters and size.

Other functionalities included are the generation of data files. In sum, it is a tool oriented to the text and articles size, also in its temporal dimension, lacking the relational characteristics of Wikipedia: categories, links and interwiki links. This means no possibility of analyzing concepts across languages or knowing how many links are directed to an article.

On the contrary, Wikiprep, another tool developed in a phD thesis, parses Wikipedia dumps to extract extended XML files adding and deleting information about relations between categories, links, and its statistics. Nevertheless, it is not a tool for examining different Wikipedia features at the same time but these concrete aspects like categorization, links or text analysis from NLP perspective.

Lastly, Hecht (2010) presented an opencode API, java and MySQL based, called WikAPIdia. Its main functionalities are compatibility for 25 languages, spatially referenced by nature and including some semantic relatedness algorithms. Hecht, gives an start-up document called Getting Started which explains computer requisites and which are the first steps to use it.

In these, the most important is a RAM memory capacity of a minimum of 4 GB, which becomes insufficient for Wikipedia editions like English in the parsing process. WikAPIdia [11] is probably the best tool for working with Wikipedia dumps since it allows managing both quantitative features like size of articles and relational ones like links.

It opens to the possibility of working with articles existing across languages, what can be called a ”Universal Concept”. Despite it is not clearly oriented to text analysis, it provides a whole range of metadata about links tags, synonyms and categories.

Still, some functionalities like edits and editors history are partially implemented and does not provide satisfactory results. Others like PageRank algorithm, which was included in a library from JUNG (Java Universal Network/Graph Framework) [12], is not fully implemented and linked to the database.

Since the more versatile is wikAPIdia which is both able to do quantitative analysis and treat Wikipedia as a relational construction it will be a base for further development and try make it complete in other aspects. One of the first procedures and next chapter theme is creating an analytical model to detect all the rellevant aspects in Wikipedia and then develope them in a tool. In table 1.1 there is a summary with all the previous tools and their features.

Tool	Project	Language	Main Characteristics	Purpose	Output
MWDumper	Csmolin Linux Community	Java	Parse MediaWiki Syntax	Import to Database	MediaWiki DB
MWImport	MWImport	Python	Parse MediaWiki Syntax	Import to Database	MediaWiki DB
XML2SQL	User Tietew	Python	Parse MediaWiki Syntax	Import to Database	MySQL DB
Pywikipediabot	MediaWiki	Python	Scripts: parsing/editing	Maintenance/Backup	XML Files / Edits
WikiXRay	phD thesis - U. Rey Juan Carlos	Python	Edits, bytes, words, etc	Quantitative	Text Files & Graphics
Wikiprep	phD thesis - Evgeniy Gabrilovich	Perl	Categories, links, text, etc.	Categories/Semantic	XML Files
wikAPIdia	phD thesis - U. Northwestern	Java	Interwiki, links, bytes, cat.	Hyperlingual, Diverse	MySQL DB

Table 1.1: Summary of tools for working with Data dumps

1.2.2. Cultural configuration

The unprecedent popularity of Wikipedia, along with its particular operating mechanisms, have attracted the interest of academics (Cindy Royal and Deepina Kapila) [18]. Rigorous studies have been deployed to know how it works and if it is done correctly. These cover topics such as reliability [6], cultural differences in the behaviour of the edit task of different nations [17], or the cited growth of internal rules [3]. The motivation in a project of this magnitude and on these conditions it stands as one of the main research problematic.

In fact, to understand who makes the repository grow and how could be the first step of any thesis. One of the major statistical studies on contributions shows how the distribution by number of edits per users of Wikipedia equated to distribution of a power law Voss [20]. That is, a small number of users are in charge of a very high amount of content. Why they do that, which are the preferred topics to be discussed and what

are the motivations leading them to invest many hours in this work, are problems to be reviewed.

At the same time the selected tool to analyze Wikipedia, wikAPIdia, provides compatibility with multiple language editions. Each of them with a different cultural backgrounds, developed around the same main topics but in different progression and specificities. The interest of this research is on how there can be a cultural influence, starting from its motivations to later see it in the configuration of the repository, topics and growth.

1.2.2.1. Motivation of editors

The movement of open source or free software (Free/Libre Open Source Software) is often used as a reference point to understand the user's motivation. In this type of projects with an ideological and contributory parallelism we can see the tendency in studies to polarize motivation as intrinsic, defined as "a fun that occurs in the course of the action" [19], or extrinsic, "to obtain tangible rewards" (Ryan and Deci, 2000). In repositories of information, other complementary explanations are adopted which have to do with ideological and social issues of the dynamics of the community or the experience of editing by the user.

Studies such as Maxwell Harper [8] demonstrated how showing the rules and data contribution from other users in film review system incremented in the number of visits and marks. Peddibhotla et al. [16], about reviews repository of the on-line store Amazon, study the behaviour from those which create the most of part of the content and the motivations that lead them to do so.

From the processing of profiles of thousands of users, it can be concluded that the main motivation of external origin was social affiliation, followed by altruism and reciprocity. In his characterization, internal source of motivation can also be discovered. These have influence in this order of self-expression, personal development, utilitarian motives and fun.

For Nov [**?**], after surveying Wikipedia editors, the main motivations are the "fun" of the activity and the "ideological" background. The survey, with 151 valid responses (92.7 percent of men), also pointed out reasons such as improved career by having new contacts, new perspectives to understand the content, feeling of being needed for writing for the community, the will to help or protection, meaning that they feel less bad about themselves just by writing.

Forte [5] exposes that the Wikipedia community moves toward a notion similar to the credibility of the scientific community, although users often are not valued in the same way. On intrinsic motivation, Xiaquan Zhang and Feng Zhu [26] show that editing any article always stimulates the previous editor to continue contributing to it.

Regarding the previous approaches, Heng-Li Yang and Cheng-Yu Lai [25] introduce two new concepts mentioning the internal images of the self. These definitions correspond on one hand to the perception of oneself and one's own standards, and on the other to the importance of a reference group. After launching a questionnaire, it can be concluded that the internal self-image affects much of the attitude of the editor.

The efficacy and self-esteem are also claimed by Timme Bisgaard [2]. His thesis, using qualitative techniques, argues that Wikipedia offers an experience of a single value, which improves self-esteem and which can all be assigned to the task you want.

In short, it can be said briefly that the research findings agree fundamentally on the existence of a right of internal origin. There are other notions of self improvement or membership in a community that you require but do not get the same degree of representation in any of the methodologies used.

However, among these common grounds one can understand that there is a common type of motivation unmentioned, suggested by informal surveys conducted by the association Amical Wikipedia of the Catalan language version, in March 2009, in October 2009 and in December 2010. The national motivation type.

Throughout questions in polls the interest in the national issue can be appreciated. In every aspect of the activity that is Wikipedia. Specifically, the latest poll asked in which topic the respondents experience or witness more conflicts and it turned out to be the first. Also, it was the second one in the number of visits (34.2%). At the time, "to promote the Catalan culture and country" is one of the main reasons why people

began to edit and therefore, issues related to the Catalan Countries (Catalan speaking area) are ones of the most chosen to write about. This new type of motivation has the clearly stated will of identifying the editors.

1.2.2.2. Topical coverage

To know the impact of a community in a language edition of Wikipedia there is no other way but to observe it at a moment in time. Not all subjects receive the same representation. Halavais and Lakaff [7] in an analysis on the thematic coverage compare the content with the publication of traditional books with regard to the Encyclopaedia Britannica and Encarta. With the election of 3000 articles in random categories from English Wikipedia edition and using the Library of Congress, two human encoders and a system for measuring (LC) which gets the length and number of articles in editions. It concludes that the strengths were in general science and history. It also highlighted the extensive coverage of popular culture, especially music bands.

The study by Aniket Kittur et al. [13] obtains a quantification and classification of the extent of the main categories in the entire content of the English version. Through an algorithm that maps as long as it descends through links of categories, they get the following results: 30% for Culture and Arts, 15% for People, 14% Geography and places, 12% Society and social sciences , 11% history and events, 9% natural and physical sciences, 4% Technology and applied sciences, 2% Religion and belief systems, 2% Health and Fitness, 1% Mathematics and logic, and 1% Philosophy and thought.

While this distribution shows the clear preference of editors, the visits give the most consulted ones for 2009, in the English version, ”Wiki”, ”The Beatles”, ”Michael Jackson”, ”Barack Obama” ”Deaths in 2009”, among others. A complementariness between pop culture icons, historical figures and technological issues. No wonder then that over a period of two years (2006-2008), there was a growth of +210% for Culture and art and 97% for People, two highly visited categories, while others such as technology, has a drop of - 6%.

Such unevenness in quantity of content generated different responses in the community and the research. Halavais and Lakaff consider it of little importance, provided that there is coverage in the encyclopedia on those fundamental issues. For co-founder Jimmy Wales it is a systemic bias to be overcome, as recognized by the Wikimania 2006 keynote [21].

1.2.2.3. Autoreferenciality

Considering a bias on the content, Brent Hecht [10] introduced the concept *Self-focus bias* in his paper *Measuring Bias in Self-focus-Maintained Community Knowledge Repositories*. As stated in the following definition, it is a bias on some sets of articles since there might be a lack of agreement on its importance.

> ”We define Self-focus bias as occuring when contributors to a knowledge repository encode information that is important and correct to them and a large proportion of contributors to the same repository, but not important and correct to contributors of similar repositories.” Hecht and Gergle (2009)

From the way Hecht (2009) implements his concept in the tests, *Self-focus bias* is primarily a semantic dimension, since the very first thing is selecting articles with a certain unity of meaning like the ones with geographical position coordinates. Secondly, those articles are selected on the condition that they must be comparable and so they must exist in other language editions. This is a second dimension of replication. And third, the study checks the prominence of the articles to know its weight related to the whole language edition (Figure 1.4).

This weight is measured by the number of incoming links (inlinks) from each article or with the numeric value that provides the PageRank algorithm, which will be described in section 2.1.2 from Chapter 2. In a wiki structure, if an article has many *inlinks* it means that it is useful to others. The PR algorithm, in the same way, uses these links to calculate the relevance to the structure of links. What you get is the relevance of geographical areas for each language edition. Hecht (2009) reveals how bilingual countries or with historical affinities are represented - for example, Quebec in the French edition of Wikipedia.

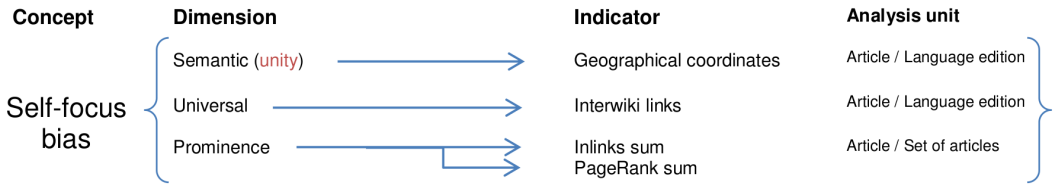

Figure 1.4: The three dimensions of Hecht (200) *Self-focus*

Once these results are obtained, to determine the degree or ratio of *Self-focusness* (SFR) of a whole language edition (WL) a relationship is expressed in the formula 1.1. It takes into account the value obtained from the territory related to the language edition (l) which produces the highest sum (C) of inlinks (then CL=l), divided by the territory non-related to the language edition with highest sum of inlinks.

As any ratio, greater than 1 means that the Wikipedia language edition arises more interest in the local region than in any other country in the world. For the English Wikipedia edition, for example, the ratio is produced by the sum of inlinks to the U.S. and France territories set of articles. Of the fifteen languages in which the test was done, tested the hypothesis *Self-Focus Bias* as confirmed.

$$SFR\left(W_{L=l}\right) = \frac{max(C_{L=l})}{max\left(C_{L\neq l}\right)} \quad (1.1)$$

It should be noted that although the ratio is made on relevant data, two points on the operationalization and its representativeness should be introduced. On the one hand, we see in the choice of two territories with more weight a bias that ignores all the other content which could meet the definition of *Self-focus bias*. It is not known if the second territory in number of inlinks (for the language edition and non-language edition) have few links less or is far. Nor in this election we can ensure that editors preference for a particular alien territory means that the content on their own territories has more or less prominence. Because of this, one can believe that it would good to use the ratio of the sum of all the local countries related to the sum of the rest.

On the other hand, the election of the relations are not strictly comparing language editions, the interest and importance attributed to a specific content as the definition shows. It takes into account the prominence of the articles located in the territory of the language edition without knowing the value represented in other editions. We do not know if the interest, e.g. United States as a the sum produced by the selected set of articles in the English version, is far from what is given by the German version.

It would therefore be appropriate to compare the relative value obtained from the sum of United States inlinks divided by total of inlinks of geographical pages in the English edition, with an average of the relative values from what United States got in other versions. This might be an option that would maintain the same specific objects of study but a different analysis.

CHAPTER 2. APPROACH

In this second chapter it is proposed an extensive technical characterization on Wikipedia and a use case defined according to the obtained characteristics. Once unveiled Wikipedia as a socio-technological object, the current tools have been presented in their main characteristics and purposes. Also the previous research on motivation and topical coverage, both necessary to understand how Wikipedia can be subjected to cultural influence.

For this, the tool wikAPIdia stands out with its multilingual compatibility and being able to manage quantitative and relational feature extractions from the repository. Nevertheless, it has some lack in particular in Wikipedia structures like for instance the history of edits and editors. It is the aim of this chapter to make a good technical characterization including new features which wikAPIdia may implement.

On last section, the use case will be presented next to the concept Autoreferentiality, which expresses the cultural configuration or influence on the knowledge repository. This concept will be expressed by several dimensions of analysis which in turn will reflect a divergence on interest by the editors of any Wikipedia language edition. Any dimension of analysis will show a particular kind of interest (in articles text, in their relations or among time) reflected on a Wikipedia structure or data characteristic.

2.1. Technical characterization and model

The main four elements are Wikimedia Foundation, the XML Data Dumps and the wikAPIdia framework (with its database) (Figure 2.1). The more in depth the content of the XML files and the previous research are understood, the better will be the model provided from the tool to be generic, scalable and useful in different scenarios and problems.

The first step of the software wikAPIdia is obtaining and treating the data from the files. Its database is filled and the framework is ready to analyse Wikipedia in one or more languages. The following pages will be based on the XML Data Dumps and the current situation of the wikAPIdia framework in order to detect what might be important to implement.

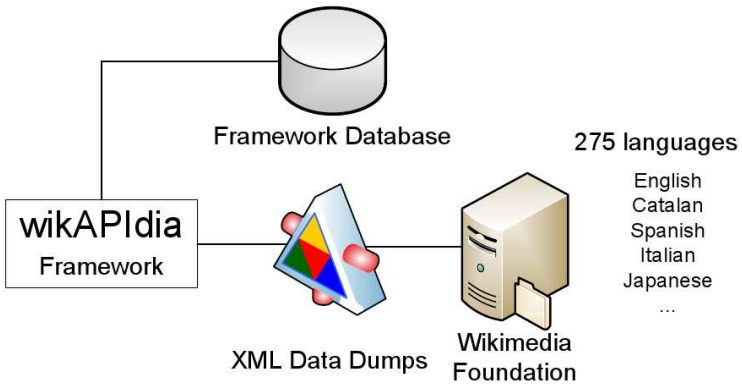

Figure 2.1: Diagram with the three main blocks

2.1.1. Structures and characteristics

As it has been seen in the table of data sources, Wikipedia's main structure is the article page. Categories, complementary, serve as a method of semantically classifying articles by topics and form a taxonomy structure ranging from general to specific, sometimes with paradoxes like a category being its own great-grandfather due to the collective editing process.

The third kind of elements is the relation articles-articles, articles-categories or categories-categories. This

is a very particular aspect of wiki structure noted before. In order to approach these three, in Figure 2.2, there is a table with the main structures and features they provide, its availability in XML files and wikAPIdia tool. In the last column, its future development is shown.

Data Structure and Features	XML	wikAPIdia	Planned
Article			
Title	Yes	Yes	=
Text	Yes	No	=
Textbox	Yes	No	=
Edits text / time / user	Yes	Incomplete	Yes
Discussion text	Yes	No	=
Temporal reference	Yes	Yes	=
Geographical reference	Yes	Yes	=
External reference	Yes	No	=
Similar pages	Yes	Yes	=
Category memberships	Yes	Yes	=
Date created	Yes	Incomplete	Yes
Images	No	No	=
Links labels	Yes	Yes	=
Category			
Title	Yes	Yes	=
Text	Yes	No	=
Category memberships	Yes	Yes	=
Relations			
Category links	Yes	Yes	=
Interwiki links	Yes	Yes	=
Links	Yes	Yes	=
Disambiguation links	Yes	Yes	=

Figure 2.2: Data structures and features: main classification

It can be seen that all article data is in the XML, less its images. WikAPIdia identifies them by their titles and gives their main features, like link labels, category memberships, temporal and geographical reference, as well as similar pages/synonymous.

Article text is omitted, as well as Textbox, both basic information for natural language processing. Nevertheless, it must be said that text is the greater part of the XML file and it would load the database with too many GigaBytes. A slow solution provided by wikAPIdia is to process every time it is wanted to recover a text the XML file in search of a specific article text by its local article identifying number.

Considering to include the text, the amount of data to import to SQL would make it unfeasible for the amount of times it would be consulted. Therefore it is a unplanned feature. Similarly, the discussion pages and text are not implemented and not planned either. It is easy to argue that it is not important information from Wikipedia, unless when it comes to quantify the activity and final result of an article. The inclusion of its text would only matter for language processing and would well be important for some problematics concerning the analysis of controversy.

XML files, due to history of edits, come with all contributions from all editors in a Wikipedia edition. This is an important feature which is currently incomplete in wikAPIdia due to an incongruent results from the parser. Each edit is marked with the whole article text, the editor who sign it and its time. This means that all the textual features like all text or derived could be developed. It is planned to include the time, user and type of user of every edit instead. These features related to an edit may ensure the analysis from a temporal perspective without overloading the database.

Category features like its title and its memberships are included. The text, which sometimes displays a summary of the category topic, is not implemented and not considered for the same previous reasons. Instead, all the different relations are included and developed in wikAPIdia.

Labels from a link are in their corresponding article or category but the relation is stored apart whether it is between Wikipedia languages, articles, categories or disambiguation pages. This last provides different articles which may respond to a word/link.

Summary Features

Since it is clear that Wikipedia is article oriented and based on categories, another way of looking at it is providing summaries of characteristics. Then, all the relations can be assimilated as inputs and outputs belonging to articles and categories. From this point of view Figure 2.3 extends derived features from the previous ones.

The main advantage of this is shortening the time of accessing this metadata, and its drawback it is a certain redundancy. Despite this, they will be implemented since the number of links is important for both incoming and outcoming (to other wikis or articles).

Total number of edits and number of editors, as well as by type (bot, ip, user), are clear activity indicators and good summary features to include. The access to edits database is a high cost procedure in terms of seconds - on a table with dozens of millions of rows. Thus eight new time related features will be implemented by wikAPIdia.

On categories, each of them could include a summary with the total number of memberships, as well as the number of articles and the number of categories. This is not implemented and planned to be, since most of the studies on Wikipedia are interested in articles.

Data Structure and Features	wikAPIdia	Planned
Article		
Num. Interwiki links	No	Yes
Num. Inlinks	Yes	=
Num. Outlinks	Yes	=
Num. Edit count	Incomplete	Yes
Num. Editor count	Incomplete	Yes
Num. Edits by type (bot, ip, user)	No	Yes
Num. Editors by type (bot, ip, user)	No	Yes
Num. Category memberships	No	Yes
Category		
Num. Memberships	No	=
Num. Articles	No	=
Num. Categories	No	=

Figure 2.3: Data structures and features: feature summaries

2.1.2. Other indicators and methodology

Having stated a simple characterization of Wikipedia, it may be useful to look at some aspects in more detail. Thus there can be created some new complex derived features to highlight some characteristics from articles or the whole repository. In Figure 2.4 there are six of them, divided into some NLP usual estimates like Semantic Relatedness or Tf-idf calculation, and other coefficients-based on links or edits features like PageRank, Edit diversity or endogamy inlinks and endogamy category memberships coefficient.

Data Structure and Features	wikAPIdia	Planned
Article		
PageRank	Incomplete	Yes
Semantic Relatedness	Yes	=
Tf-idf	Yes	=
Edit diversity coefficient	No	Yes
Endogamy inlinks coefficient	No	Yes
Endogamy cat.memb. coefficient	No	Yes

Figure 2.4: Data structures and features: calculated features

PageRank is considered to be one of the most important link analysis algorithm and which gave birth to Google Internet search engine and whose purpose is to establish the prominence of a node on a graph. It works by assigning numerical weighting to each element of a hyperlinked set of nodes - which in Wikipedia are articles - after a complex process of votes.

Each incoming link into a node counts as a vote of support in a recursive process which may need more than one iteration. A simplified algorithm is presented. The PageRank value for a page u is dependent on the PageRank values for each page v out of the set Bu (this set contains all pages linking to page u), divided by the number L(v) of all outlinks from page v.

$$PR(u) = \sum_{v \varepsilon B_u} \frac{PR(v)}{L(v)} \quad (1.2)$$

PageRank is also described as a probability distribution used to represent the likelihood that a surfer could randomly access an article by clicking on links. Often there is a complementary probability called damping factor according to the surfer would be eventually stop clicking. The whole calculation of it can be understood as a Markov chain.

Semantic Relatedness concept measures how likely are the meaning of two contents. This can be achieved for instance by defining distances between words or statistical means as a vector space model to correlate words and textual contexts. In wikAPIdia there is an implementation of it by using its outlinks, then if two articles on the same topic in different languages contain coincident links this mean they define the content on the same way.

Another text mining and information retrieval calculation is the Tf-idf weight (term frequency-inverse document frequency). This is implemented in wikAPIdia and its purpose is to know how many times a term is repeated in a text and thus its importance. It is used in some search engines as a central tool to rank documents by relevance and in Wikipedia can be applied to an article or a set of articles to see which one fits most a topic.

Edit diversity coefficient is a new calculation which wants to sum up in a value how distributed are the edits among all the editors who participated in the activity of creating an article or a set of articles. In other words, it is the tool to see if a majority of the content is created by a minority of editors. It first sets into a ranking the editors by its number of edits and then calculates how many editors are required to fill a majority percentage of edits (which could vary between 60 to 90% from the total) in relation to the total of editors who had at least contributed once.

Another two coefficients, endogamy inlinks and and endogamy category memberships are based on the idea of a set of articles and their relations. Its purpose is to know if there is semantic unity or how a topic defined by a set of articles is heterogenous or homogenous. Counting the number of links coming to a set from the same set might tell if the content is defined by the same terms. The same with its categorization.

Last, other features can be easy calculated. Most of the Wikipedia areas of interest are related to the content but also the editing process, since it obeys a complex system of rules and roles. Then it is not difficult to build tools to detect coincident editors between sets of articles or weight articles according to their location in the graph. Other coefficients may need further calculations subject to variation and then be obtained by the use of a classifiers or predictors.

Methodology

All in all, these features are important since they represent several aspects of Wikipedia graf and articles. Then, introducing them into a model to solve a social problematic may need a simple methodology in order to verify an hypothesis. For instance, the number of inlinks to an article are a clear indicator of its prominence and the number of edits goes related to the interest on the topic or how it is written.

Although each value obtained by feature is descriptive it may be related to the general trend on a Wikipedia language edition. It does not matter if an article has few memberships into categories if the categorization process in a language edition is based on one or two. Therefore an easy to apply methodology could be subtracting the average of a feature calculated in an article or a set of articles from the average of the same feature from the whole language edition articles related to this last average (1.3).

$$IndicatorValue = \frac{avg(set(f))) - avg(lang.edition(f))}{avg(lang.edition(f))} \quad (1.3)$$

If the value must be compared to others from different language editions it might be useful to make it relative by dividing it by the whole language edition articles average. Then after multiplying by a hundred the value is transformed into a percentage which is easier to read.

An important measure to apply to a feature might be a correlation with other features. To use this statistical operation with long series of articles' values in order to determine if they are alike it is important to know if they information is really new.

2.1.3. Workset

After the characterization there is a need on an entity which can permit iterating the articles to process their characteristics according to a methodology. Hence Workset is defined providing two functions, selection and storing, which permit obtaining articles and categories according to specific criterias. While the storing method may be implemented with two different options like to database or to file, the selection method will vary a lot.

For instance, a Workset may vary from another when its selection is based on a particular threshold on a feature like ”those articles which exceed a certain number of Bytes and the categories they belong to”. In Figure 2.5 it can be seen how VersionSet is that containing all the articles, RandomSet another one with articles obtained by random method, GeographicalSet with those with coordinates.

AutoreferentialitySet is the one which will be described in the case of use and wants to show the cultural configuration on language editions. Its method of obtaining articles is related to the category memberships and titles. It is important to notice that every instance from Workset may add special properties to fill up some necessities in the selection process or identification of the set.

The importance of having a Workset also lies on the hyperlingual characteristic of wikAPIdia. It makes possible to manipulate several sets of articles related to the same topic but in different languages. Then it is important to have a property language defined.

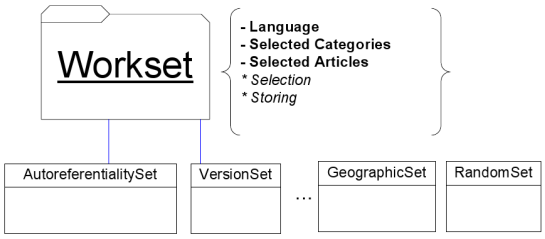

Figure 2.5: Workset properties, actions and instances

2.2. Autoreferentiality case

To start looking for a cultural influence on the Wikipedia activity it appeared that a national motivation was not proposed. Neither on topical coverage studies there was a specific analysis on the cultural manifestation in the repository. Although Self-focus bias, a concept proposed, had a certain similarity with a cultural influence it just took into account the prominence of geographical articles.

Autoreferentiality as a concept is based on Self-focus bias definition, the divergence of interest on some content across editors and similar repositories. But it wants to be focused on national or linguistic themes and it understands interest beyond prominence but on a wider range of features which will be described.

First, there will be a selection of articles to obtain the specific object ”set of representative articles” and then to obtain features converted into indicators. Later, those which are clearly validated will be used in an index, giving a definitive value for the concept also comparable among language editions.

2.2.1. Analysis dimensions

The degree of Autoreferentiality is measured using that content from a repository ”important and correct to their contributors and a large proportion of the same repository but not important and correct to contributors of similar repositories”. For Hecht Self-focus Bias had three dimensions: **semantic** dimension, **universal** or propagation and **prominence**. Here a new operalization of the concept in seven dimensions is presented.

In its procedure, Hecht selected articles which obeyed geographical classification. These represented information related to the editors and to their linguistic area compared to other areas. Although, these were not the only ones since popular culture, traditions, gastronomy, may also be included as part of the culture from the language in which is written that particular edition of Wikipedia.

Hence, the first dimension, **semantic** (1), is proposed to included the maximum number of articles with references to elements representative of the national or linguistic sphere of the editors. Then it seems a good option using its articles and categories titles, as well as their memberships to obtain a large set of articles. Introducing key words would get the first set and then crawling in their memberships might initiate an iterative process of recollection.

Hecht only selected those articles which were shared among language editions. This assumed a minimum external interest. From the current study this limits the selection of articles and it might be better to use interwiki links feature as an indicator of a second dimension of **isolation** (2) or external interest. On Figure 2.6 there is a representation on axis of external and internal interest.

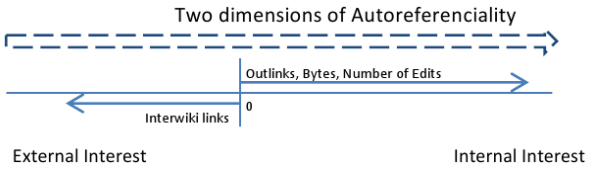

Figure 2.6: Graphical axis on the two Autoreferentiality first dimensions

New dimension internal interest or **effort** (3) accounts the attention received on articles using its number of Bytes and number of Outlinks features as indicators. Nevertheless it is unknown how many editors and edits contributed in this creation.

The dimension **prominence** (4) complements internal and external interest with a relational importance. Hecht used number of inlinks feature and PageRank value as indicators. Complementary to these, an article classified as member of many categories represent also a higher interest and prominence.

The prominence from the set into the set is the dimension called **endogamy** (5). As stated before in other indicators section it is calculated as the number of incoming relations to the set with origin in the same set

divided by all the incoming relations. More than 50% means endogamy.

The dimension **edition** (6) uses as higher interest indicators the summary features number of edits and editors. Also, a subgroup of more active editors might edit in the selected articles. The diversity coefficient will be also used in a percentage of edits 80% to see this possible divergence in interest.

To end with, **temporal** (7) dimension is defined taking into account the rate of article creation feature as interest indicator comparing the set and the whole language edition. Then it uses relative rates to their total number of articles. At the same time every period subtracted to the previous one may tell if there is an anticipate interest on the set than the whole language edition.

There are seven dimensions to analyze the autoreferentiality of a language edition. On Figure 2.7 there is a summary on the typology of dimensions, indicators and analysis unit to apply.

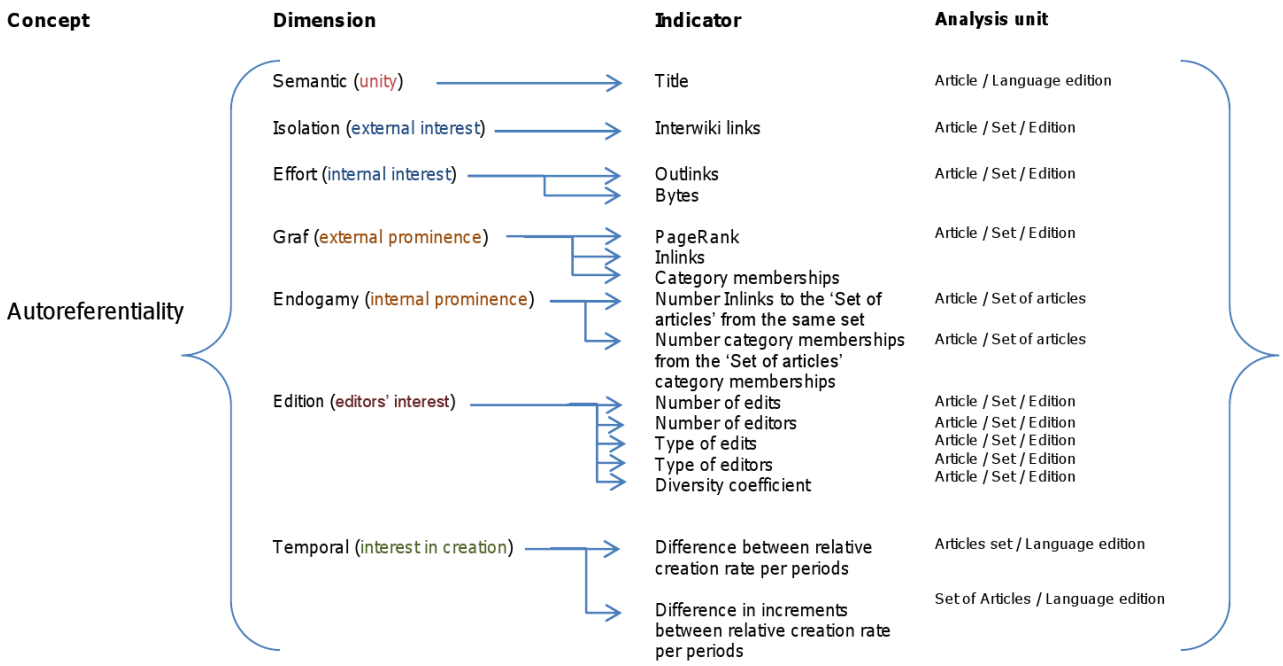

Figure 2.7: Dimensions and indicators from Autoreferentiality

2.2.1.1. Index Creation

In order to create the index to measure autoreferentiality a methodology will be applied to the indicators which will finally lead to make an adition of the most significative. The individual indicator evaluation will be the same one sketched in the previous section as a generic methodology. Also, a weight on indicators will be taken into account as well as the percenatge of the set of articles on the size of the language edition.

As said before, the procedure will first measure the distance between the feature average value from the selected set of articles and the whole language articles. Then, the indicator value will be this distance relative to the language edition average, since it is the cultural reference. Only in this case of endogamy a different rule like a minimum of 50% of links will be applied.

The weight of an indicator will be calculated as an average of all the indicator values from the tested language editions. An indicator is good generally when there are no big variations and also when there is a big distance between the set of articles and the whole language edition.

$$IndicatorWeighting(n) = \sum_{m}^{l=1} \frac{IndicatorValue(l)}{m} \quad (1.4)$$

In the end, the three factors are multiplied (indicator value, indicator weight and set of articles percentage)

for each indicator which has been tested positively the hypothesis and is not highly correlated with another one. The sum of them generate the index value. (1.5)

$$IndexValue(l) = \sum_{m}^{n=1}(indvalue(n) \cdot indpond(n) \cdot setpercentage(l)) \text{ (1.5)}$$

2.2.2. Language editions

Autoreferentiality is a property inherent to each language edition. It depends on the culture but it might be universal. In this study the choosen languages will be from the five continents, mainly European originary languages because of their Wikipedia activity.

This selection includes the Arabic (wiki language code: ar), Catalan (ca), Czech (cs), Danish (da), German (de), Guarani (gn), Finnish (fi), Hebrew (he), Hungarian (hu), Icelandic (is), Italian (it), Indonesian (id), Japanese (ja), Korean (ko), Dutch (nl), Norwegian (Bokmål) (no), Romanian (ro), Swedish (sv), Turkish (tr) and Chinese (zh).

At the following Table 2.1 there are the main language characteristics: its position in the ranking of articles, editors, active editors, activity (articles/active editors), number of speakers, number of countries where it is official and its language family. These characteristics show the variety in use of these languages and their Wikipedia activity.

Lang	Ran.	Articles	Editrs	Act. Editrs.	Activ.	M.Speak.	Speak./Art.	C.O.	Lang. Root
ar	25	143400	357000	2609	54,96	280	1952,58	26	Semitic
ca	13	309095	81515	1751	176,52	7,7	24,91	2	East romance
cs	17	188854	141192	2394	78,89	12	63,54	1	West slavic
da	23	144206	125332	1413	102,06	6	41,60	1	East scandinavian
gn	191	1373	3246	21	65,38	4,6	3350,33	3	Tupian
fi	16	263590	171606	2199	119,87	6	22,76	1	Baltic-Finnic
he	32	115111	136874	2134	53,94	8	69,49	1	Semitic
hu	18	184585	141192	2152	85,77	13	70,42	3	Finno-Ugric
is	66	30734	19654	187	164,35	0,3	9,76	1	West scandinavian
it	5	778061	606348	9248	84,13	75	96,39	7	Romance
id	23	155198	234492	1618	95,92	150	966,50	1	Malay
ja	6	736664	498692	11871	62,06	130	176,47	2	Japonic
ko	66	157184	142522	2112	74,42	78	496,23	2	Altaic
nl	10	674078	372949	5502	122,52	25	37,08	6	West germanic
no	14	292816	190180	2490	117,60	5	17,07	1	East scandinavian
ro	21	156739	176210	1463	107,14	24	153,12	3	East romance
sv	11	388598	214757	3630	107,05	10	25,73	2	East scandinavian
sw	11	21306	8702	85	250,66	50	2346,75	3	Sabaki
tr	22	155791	330398	2561	60,83	77	494,25	2	Turkic
zh	12	346104	962445	5647	61,29	1300	3756,09	3	Sinitic

Table 2.1: Wikipedia language editions and their characteristics

The selection of languages have been also made to include different sizes in amount of articles. Those larger than a million articles have not been included since they generated some problems in obtaining certain features related to History.

Besides, to give validity to the conclusions of the study there are languages of five continents and from very different sociological contexts. In number of speakers, in size and activity of the community and in cultural roots. From the biggest like chinese, multinational like arabic, to a small one like icelandic.

2.2.3. Set of articles

The set of articles chosen are the specific object in which the indicators will be measured. A well executed selection is fundamental and all the measured values will depend on it. Once the search and collection has been done, the represented coverage will permit determining the degree of autoreferentiality. The articles chosen will have a reference to the sphere of the national and cultural of the language version.

Their thematic may be heterogeneous but they are related to their territory, culture and social activity. For obtaining the articles key words will be used that refer to the territorial and political entities of the editors' language of the version and are checked with the categories and articles.

Selection by graph

About the categories and titles, the study of Nastaste and Strube (2008) [14] explores the relationships between elements of the title and articles to see if as definitions are a precise knowledge. Among the different variations observed in the value of the title (A member of Group B, elements C and D, among others) it is identified which is the dominant element and how this is reflected in its member articles. At each level of distance between articles and main categories, the semantic coherence is weakened but maintained. Some content interference appears in a very low percentage of cases.

Using these results, the recompilation will be made using keywords to search for categories and articles containing them as dominant elements. Then, taking advantage of the links, those articles and affiliated categories can be obtained, which in turn allow the process to continue on as many levels as the design categories permit. At the same time, the algorithm should avoid the duplication of articles and categories obtained by two different ways.

In the case of Viquipèdia, for example, the category "Mountains of the Catalan Pyrenees" was obtained by the word "Catalan". It contains 33 articles and 8 categories. One of them, "Mountains of Cerdanya", will include two more: "Mountains of Alta Cerdanya" and "Mountains of Lower Cerdanya". The first disposes of the articles directly, but still the second is divided into nine subcategories, accumulating a total of 95 articles. There are three levels which, according to Nastaste and Strube (2008), would contain common meanings of the "Mountains of the Catalan Pyrenees".

For the search by words it is essential to know the status of the language and its territories. If there is cohesion, the higher political entity name may be sufficient for the collection of categories that include other territorial divisions. Otherwise, all names are used so content is not lost. The same is for the use of names of peoples and places, the second type of word that is used to classify content.

Languages	Key words / Morphological bases
ar	العراق, جيبوتي, جزر, الـ قمر, مايوت, الـ بحرين, الـ بحريني, الـ جزائر, صعيد, مصر, الـ مغرب, عمان, موريتانيا, ليبيا, الـ ليبي, الـ لبنانية, لبنان, الـ كويتية, الكويت, الأردنية, الأردن, الـ عراقي, صومالي, الـ صومالية, الصومال, الـ سعودية, الـ عربية, قطر, فلسطين, الفلسطينية, الـ عماني, عمان, الـ صحراء, إرتريا, الإريترية, تشاد, زواج, الامارا, الـ تونسي, تونس, سورية, الـ سودان, الـ سوداني, الـ عيون, الـ صحراويين, الـ يمن, الـ يمني, الـ غربية
ca	Catalunya, Catala, Valencia, Mallor, Eiviss, Menorca, Menorqui, Andorra, Alguer
cs	Česk, Cestina
da	Danmark, Dansk, Slesvig-Holsten, Grønland
gn	Paraguái, Guarani, Tekove, Argentina, Avañe''ẽ
fi	Suomen, Suomi, Suomalainen, Suom
he	עִבְרִית, ישראל, המערבית הגדה, עזה, יהדות
hu	Magyar, Vajdaság, Hodos, Dobronak, Lendva
is	Íslen, Ísland
it	Italia, Vatica, Ticin, Grigion, San Marino, Sanmarinesi
id	Melayu, Timor, Indonesia
ja	日本, パラオ, アンガウル州
ko	한국어, 조선말, 한글, 조선글 , 대한민국, 한국, 연변, 한국어, 韓國語, 한국말, 조선말, 조선어, 朝鮮語, 한글, 조선글, 한반도, 韓半島, 조선반도, 朝鮮半島, 조선};
nl	Nederland, Vlaams, Vlaanderen, Vlaming, Surina, Brussel, Aruba, Zuid-Afrika
no	Norsk, Bokmål, Riksmål
ro	Român, Moldov, Voievodina, Banat, Bănățeni
sv	Sverige, Svensk, Korsnäs, Närpes, Larsmo
sw	Swahili, Tanzania, Kenya, Uganda, Shimaroe, Comorian, Shikomor, Mayotte, Msumbiji, Kongo, Burundi
tr	Türk, Kıbrıs, Prizren, Manastır
zh	中国, 中文, 香港, 台湾, 澳门, 新加坡 (Xina, Xinès, Hong Kong, Taiwan, Macao, Singapore)

Figure 2.8: Key words / morphemes used for search

All in all, to ensure the complete extraction of local content, keywords can be complemented with their root morphemes or lexemes. Thus, it obtains those formed through a morphological derivative process containing a word based on the title. Using as a basis "Catalan", Catalanist and Catalanophiles are extracted,

among others. In Czech, for example, ces- is the basis for words as Ceske (Czech language) Česká (Czech Republic) or Československá (Czechoslovakia). This saves using each of them and avoids losing any.

Note that for those languages that use pictograms, the combination becomes easier. Only the pictogram of the main land is enough, given that the creation of new words becomes more pictograms for combinatorial and therefore the first are always included in a search. Below, in Figure 2.8 with the bases used for the eight selected languages. Typically, the territory where there are more native speakers of the language adds more content to the set of selected items with their words and bases. To know what would happen if it does not include the secondary territories, it has been broken down in the following table (2.2), the percentage provided by the first four words in 8 of the choosen languages. The total sum in number of articles exceeds the final set of articles with a search all databases for redundancy of the paths mentioned.

Lleng.	Nº Art.	% Base 1	% Base 2	% Base 3	% Base 4
ca	58170	34,9	52,7	6	4,2
cs	43495	100	0	0	0
da	55619	71,9	25,6	1,7	0,7
it	115190	95	2,7	1	0,5
nl	118175	68,7	9,6	11,3	6,2
ro	49386	87,9	8,3	0	3,3
sv	164107	52,7	46,6	0	0
zh	109516	59,2	19,8	17,5	2,8

Table 2.2: Percentage of articles in the selection obtained by each keyword/base

As it can be seen in Figure 2.9 in the Catalan Wikipedia 87.6 % of content comes from Catalonia. Considering the demographic and geographic, only 51.55 % of the population resides in the Catalan-speaking territory that occupies the 45.22 % of the total geographical area of the Catalan-speaking area. This inequality can only be explained from the answers given by the third survey by Amical Wikipedia, where 77.15 % of the respondents (readers and editors) are from Catalonia.

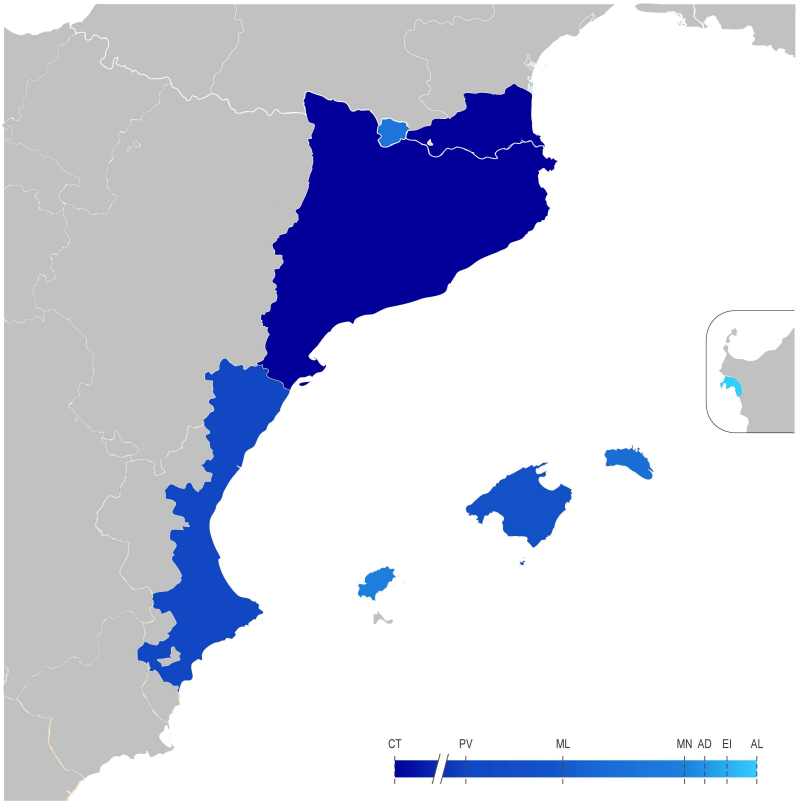

Figure 2.9: Map of the Catalan-speaking territories in the number of articles

Levels

Lang.	0	1	2	3	4	5	6	7	8	9	10
ar	2855	12177	7020	9372	8397	8710	1935	547	232	111	0
ca	1756	15621	12489	12547	4266	1963	1473	563	907	0	0
cs	2007	16102	19567	9590	1722	638	409	266	21	109	54
da	1257	26992	10106	5592	9497	14390	20656	22709	17276	4091	628
fi	2327	37031	12948	3418	686	80	98	12	0	0	0
gn	43	261	153	69	0	0	0	0	0	0	0
he	1571	18602	7837	3745	1436	686	233	29	0	0	0
hu	2430	29584	7701	3213	2814	199	42	9	0	0	0
id	1299	11551	4565	813	600	132	0	0	0	0	0
is	693	8054	1409	278	21	0	0	0	0	0	0
it	5568	34754	38541	36466	11276	4734	6365	1683	713	995	392
ja	7593	190899	125557	59672	18214	7919	8665	1132	823	813	5946
ko	2403	26384	9478	2639	987	532	114	96	31	8	0
nl	4953	59107	26055	6298	2471	2850	1749	855	158	0	0
no	1435	33699	12750	9177	9252	3758	2819	4492	395	220	0
ro	1863	27256	17360	1824	273	148	16	0	0	0	0
sv	3690	73794	19061	10687	2158,00	3409	1757	537	278	132	94
sw	258	1622	1118	2063	22,00	1	0	0	0	0	0
tr	2453	15547	6308	6054	2246,00	1685	1624	2614	3932	1241	1239
zh	4803	30364	25849	24425	9612,00	4734	2673	4651	7720	6565	2275

Table 2.3: Number of articles obtained by each level of the selection

As explained above, the descent through memberships in categories is what allows the collection of articles and new categories. The point at which this taxonomic classification expires depends on the language edition, on how the editors have decided to conceptualize the categories. In Table 2.3 it can be seen the volume of articles obtained for each level, in which zero are those containing the keywords.

A high percentage of articles is collected in the first three levels after the initial search. At the same time, level to level, the semantic relationship becomes weaker. In a qualitative test for the version in Catalan, Italian and simplified English (used only for testing) interference of content have been seen in the fifth level regarding the initial.

Each version makes the process of categorization without a centrally planned will and therefore there is always diversity in the number of memberships per article and a certain disorder when establishing hierarchies among categories. Because it is outside the scope to make a thorough check for all language versions to recognize that content which interferes, such an arbitrarily measure is taken as good for the first four levels of each selection process.

These first four levels (counting from zero, with the keywords, to the third) are called the ”actual” levels in contrast to the levels of ”selection”, which cover the ten. The actual levels will be considered for testing the hypothesis previously described as a series of self representative articles.

An alternative method for selecting the articles which refer to this semantic national knowledge would be using a frequency term calculation method like tf-idf. The same keywords which were selected to start the process with the articles and categories might be analysed in the articles to see in which they appear more frequently and what relevance they have. However, doing this alternative selection and compare the extension it gives remains as a future line of research.

2.2.4. Hypothesis

Once the concept has been operationalized in dimensions, some estimates can be made on the signs which the indicators take. If it is confirmed, it will be valid to represent the differences of the self representative articles with all the articles of the version. Ultimately, these indicators are those which allow the calculation of the index to measure a property of autoreferentiality of a version.

The hypotheses will be split in the same dimensions. At the same time, given the same data, more information can be provided on how to develop a set of items without directly using the final index. In these cases

the hypothesis is formulated and its results are interpreted alone.

The first dimension, semantic, is the basis for hypotheses. The autoreferentiality property itself is a hypothesis about the existence of common features on the set of selected self representative articles. At the same time, it is also believed that under the semantic value of content each hypothesis will manifest itself in the different levels of selection, especially more affirmative at zero or nucleus, which contains keywords in titles and is the core meaning.

Therefore, external interest (H1), domestic interest (H2), prominence (H3), publishing (H4) and the temporal evolution (H5) are assessed.

2.2.4.1. H1. Isolation (external interest)

The hypothesis of external interest used the indicator number of Interwiki links. By definition, the content obtained will raise less external interest or will be attributed less importance in other repositories. That means that the average number of all articles of all self representative articles will have less links to other versions.

H1. The set of self representative articles will have less interest in other versions, this will be reflected in an average number of interwiki links per article inferior to the total number of articles obtained for a language version.

2.2.4.2. H2. Effort (internal interest)

About the dimension of the effort, aspects such as length and number of Outlinks are measured. There are indicators which can be understood as the internal interest attributed by its own community.

H2. The set of self representative articles receive a higher internal interest which will be reflected in an average of Bytes and Outlinks than obtained by all article of a language version

2.2.4.3. H3. Prominence (relational interest) - External (Graf) and Internal (Endogamy)

The two dimensions of prominence take into account the indicators inlinks and the category memberships. The external prominence is measured with the number of Inlinks, the number of category memberships and the PageRank value, while the internal or endogamy revises the origin of the inlinks and category memberships to see if they are coming from the same set of self representative articles.

The hypothesis is that the indicators of internal prominence will have an average value greater than that obtained with the full version of articles. For indicators that represent the endogamy a favourable hypothesis is estimated, which is represented in exceeding in more than 50 % the number of relations towards itself. Also, that would confirm the semantic unity of the whole.

H3. The set of self-representative articles has more prominence as a graph showing interest in membership and relational categories. At the same time, they also relate more with themselves than with the other provisions of the language version.

2.2.4.4. H4. Edition (editors' interest)

The dimension of edition wants to respond to the activity by which articles are accomplished. Various scenarios will be presented considering the indicator. The first two, the number of edits and editors, to understand that increases with the interest the article. The estimate is that in general there will be more editors per article for the whole self-representative articles that the whole language version comprises.

The relationships between editors and editions are measured with the coefficient of diversity indicator. The maximum value, one, shows that a specific percentage of edits it took all the editors who contributed. It

can be understood that the divergence of interest among editors about the self representative content of the whole shows less diversity of the edition than in an average article of the version.

An hypothesis about the low diversity can be made. To apply it to the level of the article, it is taken by percentage (majority) 80 % of the edits of articles. To apply it to the sets (collection of self-representative articles and the version) the value for all rates [0,100] periods of 2.5 is calculated and finally the differences obtained.

H4. The sets of self-representative articles will be made in average by less editors than the total of articles of a version, but contains more bots, more registered users, and less anonymous. The diversity of the edition with what is done by editing the content will lower the overall results of self-representative articles of the language version in both mid-level articles as the curve of values of the sets.

2.2.4.5. H5. Temporal (creation interest)

On the last dimension, time, the analysis concentrates to see if the interest in creating the set of articles is superior to all the language edition and also if it is preceding in time. The rate of created articles per month will be used. The hypothesis for the two indicators is that at least it behaves like all the articles.

H5. The set of articles will grow in the same periods that makes the total language version of the article. Even at these points exceeds the relative value of number of items created on the set and also show a differential increase between periods greater than 50 % of the periods, ie, progress will be made in time to trend.

CHAPTER 3. DESIGN AND IMPLEMENTATION

In this third chapter we will present an implementation on the wikAPIdia framework of the new features explained in the characterization. For this it will be necessary to recognise the design requirements in order to provide a robust tool and versatile to all the scenarios. wikAPIdia will be detailed in its functioning of the main classes and also the database.

After that, considering the complexity of managing the amount of information of this knowledge repository as well as being one of the most difficult problems, we will briefly explain the used scenarios and its performance. Last, simplifcations on the programming files as well as the launching methods will be presented as automatized methods.

3.1. Requirements

First of all it is important to recognise the importance in design of several aspects which a new version of wikAPIdia or a similar purpose software should take into account. These are regarding good programming as well as the particular problematic of analysing and treating large data sets. When applying them the further task of testing might be easier and more friendly.

- Scalability: essential in the abstraction of classes as well as in the design of the database. wikAPIdia can still implement many other Wikipedia data structures.
- Optimization of on-fly processes: every process which runs once through a data can be optimized to obtain several features simultaneously and then save time.
- Optimization in space: when dealing with large files which can size up to 6 TB it is essential to find an alternative method to work with them sequentially.
- Duplicity: it is essential to have several copies of the most valuable selections of data in case the database collapses or for optimizing access time.
- Test automatization: when the number of tests exceeds a number for a large number of objects which to be applied, some mechanisms of partial or full automatization might be useful.

3.2. Framework

wikAPIdia is a framework initially developed by Brent Hecht from University of Northwestern, Illinois, in the course of a phD studies. It has been used in several publications and the proposed extension development will make its 0.3 version. As explained before, it is Java-based and MySQL for its database and its main characteristics are being hyperlingual, spatially reference tags for its articles and support semantic relatedness algorithms.

As stated in technical characterization, wikAPIdia provides mechanisms and tools for dealing with most of the Wikipedia data. Nevertheless, those which we considered in interest to be included require a revision of its structure, abstraction in classes and tests mechanisms. wikAPIdia has been primarily used in some particular features and languages, it is hence essential to make it more robust for longer data lengths and iterations. All the code can be found in the attached CD.

3.2.1. Main classes

First of all, we will present its three main four types of packages and classes: wikAPIdia, WikipediaDatabase, languages and database entities. This is the core of the API, mainly structured in a central class with its same API name which dispose of a database connection and have access to multiple Wikipedia structures entities and functions to create or operate with them.

Every operation will start creating an object instance of Wikapidia, which will define whose database name is it linked to and the folder where XML dumps are located. All the other classes and scripts will run through the database, then it is the most important class of the API. It connects to MySQL using the JDBC connector and can work in local and remotely.

In the following Figure 3.1 there is a diagram which graphically explains the relation between these two elements. After, an experienced programmer may use or create functions which call DBentities to operate with them. Those can be run only once, like the parsing methods or reference space tagging (spatial, time...), or can be iterative like the ones used in Wikipedia analysis.

It is essential to note that there is a package called Languages which includes a primary class called Language with its main parameters and its subsequent different editions like English, Spanish or Japanese. In the current version, up to 25 languages, but extended up to 33 (Aragonese, Basque, Esperanto, Guarani, Icelandic, Macedonian and Occitan). Every language file defines the XML tags which will be used by Wikapidia package parsing classes to import the dump into the database.

DBentities are defined by a main class WikapidiaDatabaseEntity which is identified by a language and sets the methods to obtain and store the entity. Hence all the other structures will define different parameters according to its relations and data. Like defined in the technical characterization, LocalArticle will include a title, number of inlinks, number of outlinks, etc. And a link will be defined by its start article id and end article id and their languages.

Any new extension to wikAPIdia will be adapted to this structure. For instance, other packages like SR (SemanticRelatedness) or Research provide methods to extract metadata from the text or articles relations. For them, it is necessary to create an instance of wikAPIdia class and its related database and access the database using WikipediaDatabase.

In our case, we will require changing the DBentities to acommodate new different types of information like summaries and other calculated features as well as expanding the WikipediaDatabase class with multiple functions for extracting and introducing sets of DBentities according to precise parameters or restrictions.

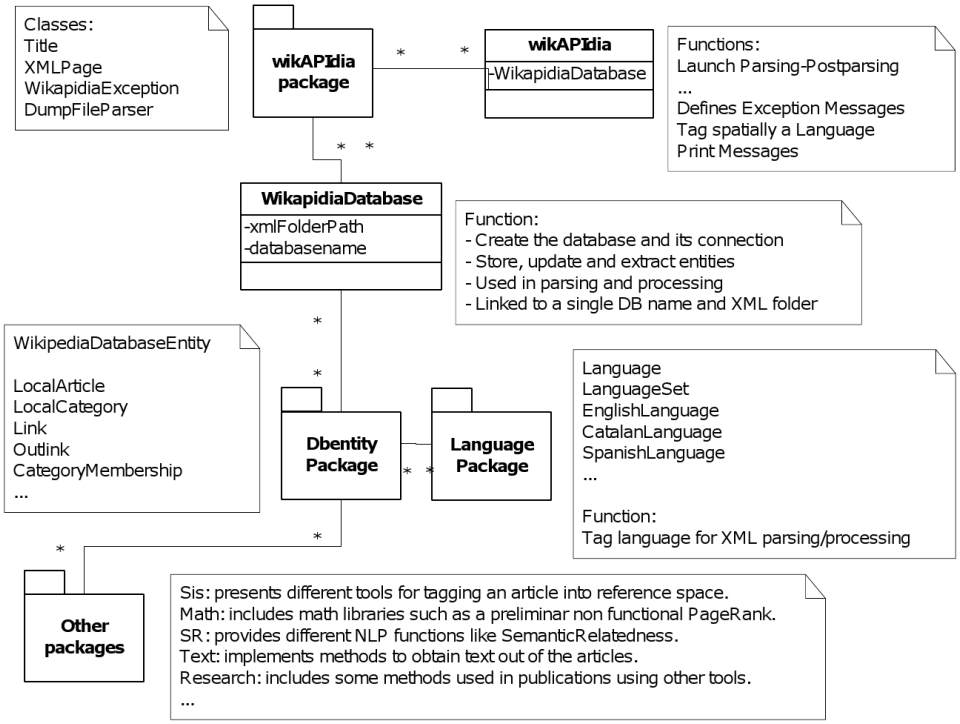

Figure 3.1: Class structure of wikAPIdia

3.2.2. Database

Wikapidia database is a MySQL database. In our experiments it is running in the LSI department of UPC (mysql-rdlab.lsi.upc.edu). It is implemented with MyISAM (while in our local version is InnoDB) and its webserver is Apache/2.2.14 (Ubuntu) with a client MySQL 5.1.41. Its indexes are B-tree. In its current implementation has 29 tables, two of which are developed oriented to periodical results.

The most loaded tables for the 25 languages which have been imported are links (382.347.410 rows - 29,2 GB), edits (201.123.346 - 15,6 GB), anchor texts (96.459.169 - 9 GB), local articles (12.384.481 - 1,8 GB) and category memberships (34.054.528 - 1,4 GB).

Initially it had 24 tables. Now, to fulfill the needs from working on a set of articles in various problematics it has been necessary to provide four new tables: local articles fromset, local categories fromset, editors fromset, results setindicators, results setlevels. Two of them store the selected entities, which are updated when their properties change as well as when some special features are introduced on them (for instance, endogamy features). Editorsfromset is a table which ranks all the editors which have participated in the selected set by their number of edits. Last, both results tables are updated once every processing technique is applied to the set.

WikapidiaDatabase new functions

As stated before, WikapidiaDatabase is the most important class from the API and it is also reflected on its number of methods (136), 40 of them are created in this project. First of all, there those which are used for creating the database, preparing its syntaxis and filling their tables once. These have been updated to the new requirements and to fit the changes in the entities (Local Articles new features, etc.).

After that, there are some others which can be used several times as the experiments require them. Their purpose is not to process the information but select or introduce one in particular, although in some cases they require searches and some simple operations.

Here we list some of the new methods in three different groups. The first would be related to storing data and results into the database. The second would be basic for processing - provides the articles depending on the needs - and the latter obtains features from the articles. The complexity of the functions from the third group is clearly higher than the second as they usually need multiple queries and calculations.

- Store articles in the set or some of their features: storeLocalArticleFromSet, storeLocalArticleEditInformation, storeEditorsCountFromArticle, writePageRanktable, storeindividualDiversityCoefficient, storeInlinkstoArticleFromSet, storeLevelsIndicator, storeIndicators.
- Obtain articles and categories according to features: getArticlesbyWords, getAllArticlesWithoutSet, getRandomArticlesSet, getArticlesbyId, getLocalArticlesLevelsFromSet, getLocalArticlesOfReferenceSystem or getTopNInlinkedArticles.
- Obtain features for processing: getnumberEditsbyTypebyArticle, getEditorsRankEditsbyArticle, getEditorsRankEditbySet, getDiversityCoefficientSet, getDiversityCoefficientVersion, getEditUsersbyTypeinArticle, getMostActiveCoincidentEditors, printMostActiveUsers, getInlinkstoArticle, getInlinkstoArticleFromSetDB, storeCategoryMembershipsFromArticleFromSet, getPageRankValue, getnumberILLS, printRanksInlinksFromSetToSettoFile.

Large file problematic

Despite that, the main problem of the history files processing was its large size which could be from 1 GB the smallest to 6 TB the largest. This difference in space due to the difference in number of articles needed to be taking into account when designing the algorithm (worst case scenario).

Also, wikAPIdia is single-thread based and could not make use multithreading techniques which require data replication for several nodes in a high-performance cluster techniques. Then only part of the data is processed and temporally stored on volatile memory at the same time. This opened a door to piping a compressed file extraction to the parsing process.

For that, Wikimedia Foundation provides the XML files in .bz2 and .7z compressed versions. The biggest file is then 300GB which is affordable for most of the current servers. When implementing the decompression into the algorithm the streams need to be redefined into a different kind of data. Decompression process is carried out by the java process and does not produce any output file.

3.3. Implementations

3.3.1. New features

Summaries

One of the particularities of wikAPIdia is that it provides all the XML dump data into the database despites text. Then some information is sorted as relations between articles when it could also be stored as a summary. Besides links are necessary to be stored when using algorithms which calculates graph properties (p.e. PageRank) in most of the cases they are not useful.

For this, we considered interesting to dispose some existing and new information in the same articles or category tables. These two are the main data structures. Articles lack the number of interwiki links and category memberships as well as all the edit information.

According to this, two solutions have been implemented: first, retrieving these features from their tables (ills, categorylinks) and examine where they point to and create a summary for each article, and second, to process and include these features into the articles table at the same parsing process. The last articles parsing process is usually the first and always previous to history parsing.

It is important to notice that in a long processes like this adding new operations may slow it. Nevertheless, increasing the time in the parsing does not make wikAPIdia parsing time much longer than other tools presented previously and it is a save in amount of time for good features like these two.

Also, obtaining these summary features on the fly require minor changes like counters per article, new columns in the database as well as a redefinition in the MySQL syntaxis message.

History features

wikAPIdia provides an initial implementation of the history data dump which imports into the database table ”edits” and ”editors” (with number of edits) as well as ”edit count” per article. The algorithm is based on the same XML tag detection patterns but have a different way of sorting the information.

The history file includes all the contributions, each signed by an author (which is either bot, ip or registered user) at a particular time. Then the DBentity will relate every edit to an editor name, a time and the article in which is made.

The main problem of the former algorithm was that it could obtain edits but it showed incongruence in the number saved into the database, the sum assigned to each editor and the count per article. This was due to a bad identification of the tag contribution and several errors in the editor database structures.

It was solved by refinating the tagging in the XML and several try and error with small XML files whose tags could also be detected and verified by text editors manually. Besides, it was important to note that the edits were better stored at the same time than obtained because of the temporal amount of memory they would occupy.

At the same time, these changes brought some awareness on the importance of the type of edits per article. The editing and temporal Wikipedia features are the least treated by the current research. Then an interesting new implementation was including the number of edits per type and number of editors per type in every article. This was achieved similary by the use of counters.

Another reason for including these summary features was for crash preventing. Opposite to other tables, the edits table is very large and is busy for long time when it is update. It needs then to modify its indexes and

adapt to the new rows. In the testing it has been necessary to use the repair. Separing the edits table from the article table giving summary features to the last one made possible to update one table and consult the other with no problem.

Slow access and indexes

The last feature we considered ”diversity coefficient” was also finally implemented to be obtained at the same historyparsing process. It is value calculated for each article taking into account a majority number of edits generated by a minority of editors. For this the choosen implementation accesses a method which counts the edits per editor by each type and introducing the percentage of edits calculated how many editors are required to reach this amount of edits.

Processing this coefficient on the fly made also enshortened the time. Otherwise, it would have been necessary to access the ’edits’ table and calculate how many edits are associated to every editor which is more expensive due to the number of rows of the table (up to hundreds of millions). After redefining the database using several index types like ’hash’ or ’b-tree’ we reached the conclusion it would still be faster to use on-the-fly techniques.

PageRank

In the case of the PageRank feature it has been necessary to rewrite the way in which the library JUNG obtained the links and identify them with an article ID. Not all the links were introduced and some of them were associated to a different kind of ID. PageRank process starts loading all the links from the database to later produce a value for each ID which is temporary stored in a HashMap. After that, there is an updating of the value within each local article row in the database.

Set related features

Last, two new features have been proposed related to the selection of a set of articles. These were presented in the previous chapter, inlinks/categories coming from the set (endogamy) and group diversity coefficient. The first is calculated throughout an iteration of link examining to see if they have an article from the set as an start and it gives a single number result. Despite it finally has not been used, there is a coefficient planned and developed with the purpose of describing this endogamy. It uses the levels from which they come and weights accordingly.

The second one, diversity coefficient for a group, is an interesting feature that can be calculated for all the percentages. In history parsing it has been explained the method by which it can be obtained. The advantage it gives over an average of articles diversity coefficient is that it can gives a more accurate behaviour explanation of the whole community. When obtaining all the results one can see a curbe which assimilates to the law power first observed by Voss [20].

3.3.2. Workset package

The Workset package wants to fulfill the requirements set in the technical characterization for wikAPIdia to work with sets of articles. The framework is complete in most of data structures and follows a logical way of storing and summarizing metadata features. Nevertheless it lacks providing a good abstraction to experiment with the information in applied problems.

For this, the Workset is an object with a Wikapidia and a language, whose main functions are selecting a particular set of articles/categories and storing them in two possible ways. Also, Workset is provided with an ID which is the date in which it is instanced. This is an important characteristic since we may want to repeat the same experiments in a language edition of Wikipedia and it is necessary to identify them in the database.

In Figure 3.2 there is a whole diagram of the Workset package, which includes all the derivated classes like VersionSet, GeographicSet, among others. AutoreferentialitySet is the one in which our study is based and it adds some special characteristics regarding its use like searchWords. Indexes are necessary for locating articles in their levels of selection within a set. EffectiveArticles or categories are the three first levels.

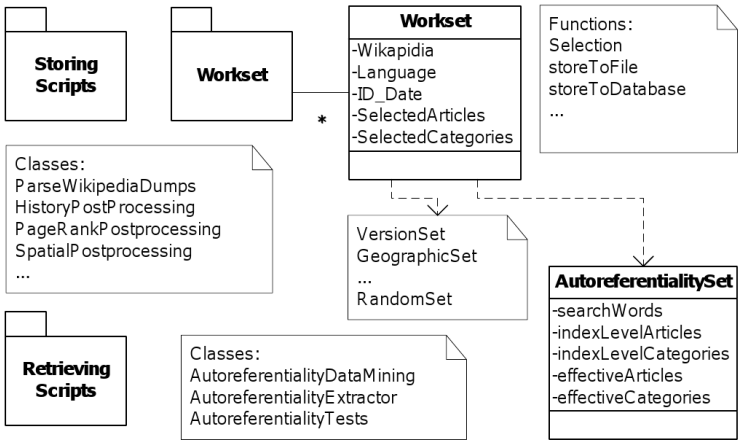

Figure 3.2: New packages and classes related to Workset

3.3.2.1. Selection, replication and consistency

Mostly, every workset is differentiated by its selection method. We have seen in the database methods that there are different ways of retrieving articles. Hence different Sets can be created for different purposes. This could be according to their reference system (GeographicSet or TemporalSet), random with restrictions or even the whole language edition (VersionSet).

Any Wikapidia database can hold one selection of articles per language at the same time in its table ”local articles from set”. Although using other storing methods it is possible to load various sets, storing a selection has sense for reducing access time to a big table like ”local articles” and for creating features related to the set. Another reason for Database storing is for external use of results, like for example presenting them on a webpage. In appendix B we present a wiki which includes all this document.

Other storing methods like serialization and to file are implemented mainly for practical reasons like not having access to database or crash. When the selection is made all the articles have no PageRank neither history features. Hence to avoid having inconsistency it is important to repeat the process after every update.

3.3.3. Storing and retrieving scripts

In the diagram there were also two other packages. One of them, storing scripts, is a ready-for-use group of classes to import and prepare the Wikipedia data for processing. For that, there is a logical sense in starting for ParseWikipediaDumps, then PageRankPostProcessing and after that HistoryPostProcessing (in between, if we want set related history features we may need the selection).

The other one, retrieving scripts provides three classes for dealing with the problematics of Autoreferentiality. Once the set is defined and instanced it is necessary to start calling the methods to process the data and later extract it. AutoreferentialityTests has this purpose with all the different tests organized to be called from the outside, locally or remotely.

AutoreferentialityTests class is a test manager and has been provided with argument input for external use. Since the most probable scenario for processing Wikipedia is not in the same IDE it is important to provide mechanisms for a .jar to run different tests or operations without having to compile it twice (consequently, send it twice too). For that, AutoreferentialityTests can be introduced the number of the test to run and the languages where it must be applied.

All the results obtained throughout AutoreferentialityTests can be printed into screen, import results to database and store into a .txt file. This last option is the one we used most since every test updates the file, which in turn serves and updates an Excel file with different sheets per dimension.

Later, all the other scripts, storing and retrieving, have been following the same syntaxis in arguments. It is something which prior was not in the design but have become so useful in the context of 25 languages and more than 17 different features to obtain per each.

AutoreferentialityExtractor is a class which instances an AutoreferentialitySet and provides many different options to extract series of data for statistical analaysis. This has been so useful when analysing articles features: correlation or probability distribution[1]. Similarly, by arguments we can call the set we want, features and number of articles. Also the kind of normalization which should be applied (from the same set, from all the articles from the language or all languages).

AutoreferentialityDataMining presents a set of methods to extract the articles and their features in the syntaxis of Weka Data Mining set of tools (classifiers and clusterers)[2]. It lies on the future work to present some results on how articles can be identified by its features. This class presents an automatized method for choosing the set of articles, their features, and how they should appear in the input file for the classifier. It has also defined functions for loading, training and testing data in a preliminar version.

3.3.4. SetProcessing tools

Last, the SetProcessing Tools class has the purpose of processing the data retrieved by the worksets through WikipediaDatabase. It provides a collection of methods for creating general statistics, sorting and weighting features such as creation date, number of interwiki links. This is where any methodological solution is applied.

This class can be divided into handy algorithms (like substracting elements from one Vector to another) or those used for processing and treating features. From the thirteen methods implemented some of them will be developed below in explanation:

- getGeneralStatsbyAttribute is a method whose input is a set of articles and a feature and returns maximum value, minimum, average and sum. It is the most use for the Autoreferentiality test.
- getTemporalStats, getTemporalStatsPercentage, getTemporalStatsCompared. These three are methods which help analyzing temporally Wikipedia. They sort articles in periods of time and then analyse their features using normalizations and different kinds of comparisons.
- getInterwikiRanks. It is an interesting method which gives as result a ranking which points where the articles are replicated (where the interwiki links are directed).
- getEndogamySetInlinksCoefficient. It is a deprecated feature which was based on the inlinks from set feature. Depending on the levels which pointed to the articles from the set the coefficient was higher or lower.
- getTopNArticles. Using different weightings for each feature we can obtain the Top N articles most rated. It is interesting for seeing articles with similar features or most representative from a dimension.

3.4. Infrastructure

In this section there is an evaluation on the different scenarios in which Wikapidia has been tested. Since there have been problems derivated to the high consume in some resources there is a cost evaluation for different scale of processes and files. This is an important section since the future user of wikAPIdia must be careful in selecting the scenario and the languages to process in order to avoid bottlenecks.

[1] An example of the Matlab scripts which have been used to correlate features in Annex X

[2] Weka, Open Source Machine Learning Software in Java - *http://www.cs.waikato.ac.nz/ml/weka/*

3.4.1. Initial scenarios

Initially, we considered using a laptop Samsung R530 with an Intel Core 2 Duo T6600 (2.20 Ghz, 800 Mhz, 2MB) Intel Pentium T4400 (2.20Ghz, 800Mhz, 1MB) with 4 GB (DDR3 / 2 GB x 2) as RAM memory. Languages so different in amount of articles like Simple English (70K articles), Catalan (300K) and Italian (600K) were working fine in the parsing for this scenario.

Considering the first bottleneck, space, 320 GB of laptop hard disk were enough to hold any compressed Wikipedia history file (english is 300 GB). Besides, the problem might be when increasing the number of language versions as well as the downloading. Second bottleneck, RAM appeared in the process of parsing the history files and PageRank, as well as some database iterations like ”inlinks from set” feature.

Apparently, CPU was enough for processing complex process PageRank but the RAM could not store temporarily the amount of links which uses. Having discarded this scenario the laptop have been used in programming and debugging in the Eclipse IDE, where small languages like Guarani from Paraguay could be processed in matter of minutes in all parsings and calculations.

The second scenario, a server from UPC with high space availability, was not suitable either. While it could store up to 4 TB, which is more than what is required for most of uncompressed history files, english one still exceeded with 6 TB. Although, RAM memory was still insufficient with 4 GB, and big languages up to a million articles can require more than 100 GB for certain processing. The third scenario, a cluster, seemed more appropiate.

3.4.2. Cluster

The presented cluster scenario called eixam is located in UPC department LC-LSI (Laboratori de Càlcul del Departament de Llenguatges i Sistemes Informàtics), parallel to other clusters like nozomi, storage and tenada. It is based on the system Sun Grid which can enqueue processes with different parameters of functioning. Every user has access via ssh to a central node from which to store the java .jar files, write the scripts and eventually send the processes.

Eixam has 18 hosts up, a total disk of 4.4 TB, and is used for processing purposes. A majority of nodes with 32 GB of RAM and one with 64. Users have access to it via ssh through two different systems depending on the architecture which they want to run, 32 and 64 bits. Then it is necesssary to set the path from where they send the processes. Any process is a script with executable permission specifying the files to execute and its location.

All of them can be monitorized using the command line via ssh in the central node either using Ganglia Grid Report web interface. Some of the flags for enqueing processes are related to the output files, to the memory reservation or the limit of time after which a process must be stopped.

Processes (and their files) are sent automaticly to the node which fits the set requirements or otherwise to the queue. It is also possible to send a process to a selected node. In our case, we used a script for every different process file: parsing, history parsing, PageRank, dataminingextraction, extraction and autoreferentiality tests.

As commented before, the user can modify the scripts introducing the arguments which will define the languages and tests the processes will run. For example, the autoreferentialitytests is called six times in one script (alltests.sh) with the tests of all the dimensions. In our cost evaluation we monitorized the four following languages: German, Japanese, Catalan and Icelandic. And the processes parsing, history parsing and PageRank. They provide a good representation of different scale cost and are essential to obtain all the features. However, as it is known the most usual bottleneck will be the RAM memory.

Clock

Any test to run with Wikipedia sets of articles can be repeated with newer versions of the same articles. For that we considered in the implementation having an ID (date) to a Workset which is introduced at the same time with the results to the database. Furthermore, the cluster scenario working with scripts enhances using

a clock.

This means that all can be fully automatized in four different tasks every expected amount of time. First, using a script to delete the XML/bz2 Wikipedia files and download the new ones. Second, connecting to the database to delete all the tables but the results. Third, running the scripts to import data to the database. And fourth, running the new tests.

3.4.3. Cost

Parser process is stable collecting data from the XML and storing part of it at the same time to the database. At the end, the other data is also stored and released from temporal memory. In Table 3.1 we can see a very important difference on the estimated consumed RAM from a 5,4 GB file to a 8,4 GB due to that reason. In this first process we can conclude that processing languages higher than 600000 articles like japanese might find a bottleneck.

In other details, the average rate in processed articles does not depend on the size. Neither the CPU time compared to the real time (indicated as clock), since it is related to the free CPU time the node has (since they all run other processes). Since we could not obtain the real RAM used by the parser process we had to calculate it through the graphical monitoring like in Figure 3.3, despite there was also some other RAM consumed for the same reason. Memory counted as GB·CPU is the multiplication between the CPU time multiplied by 2,9 GB.

Parser	File Size (GB)	Node	RAM	CPU	time wclock	time CPU	estim. RAM	mem GB·CPU	Avg. rate
de	8,4	111	66103024 KB	2926 MHz	5,96 h	2,46 h	58 GB	32200,86	75,12 articles/sec
ja	5,4	112	33019548 KB	2992 MHz	3,56 h	2,67 h	4 GB	29819,742	104.66 articles/sec
ca	1,4	117	33009420 KB	2660 MHz	1,35 h	22,4 min	3,5 GB	4082,697	75,11 articles/sec
is	0,131	106	33009420 KB	2659 MHz	9,81 min	1,83 min	1 GB	336,272	100.1 articles/sec

Table 3.1: Main cost parameters of parsing

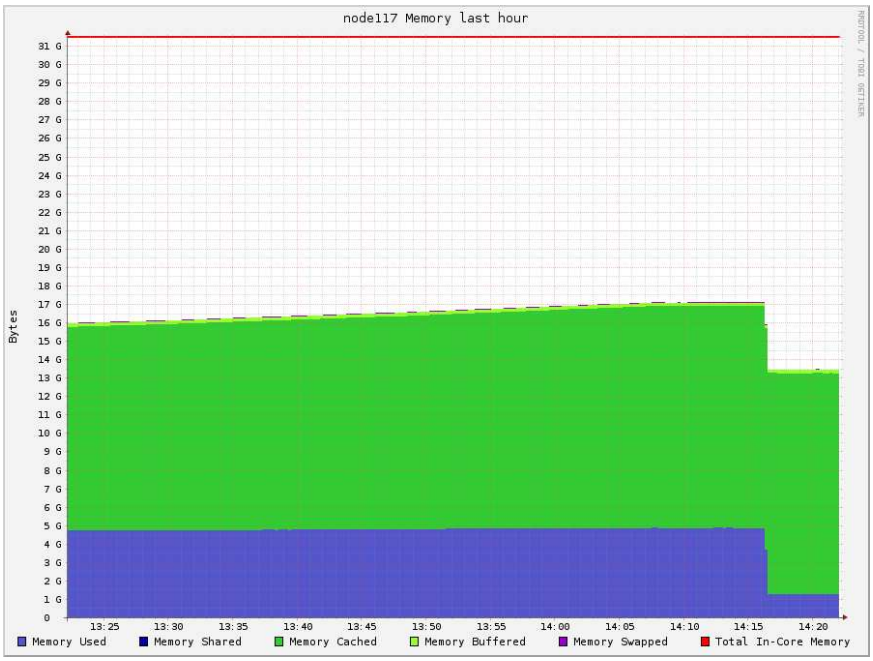

Figure 3.3: Memory consume for the catalan language parser process

Overall, in the parsing the amount of data transferred was at a ratio of 0,6 MB/sec input to the node from eixam to 0,5 MB/sec output.

On the second Table 3.2 with History Parser process we can see how the RAM consuming overpassed the 4 GB for the catalan language edition file. While its size it is clearly much higher (even compressed, uncompressed is up to 11 times the compression size), the amount of time (clock) spent parsing is proportional to it. The ratio with CPU time is still at low performance due to the other processes.

In Figure 3.4 there is a graphic with the memory occupied by the history parser in german language edition and its network use. The temporal storage is stable while the network use is oscillating due to the cycles in the algorithm where there is some temporal storage and then introducing to the database. The little rate in output is due to the few numbers (diversity coefficient, edits count, etc.) it produces.

Hist. Parser	F.Size (GB bz2)	Node	RAM	CPU	t. wclock	t. CPU	est. RAM	mem GB·CPU	Avg. rate
de	94,7	111	66103024 KB	2926 MHz	49,409 h	34,581 h	46 GB	53283,242	104.66 art./sec
ja	21,5	117	33009420 KB	2660 MHz	25,468 h	9,881 h	15 GB	31487,889	104.66 art./sec
ca	2,4	117	33009420 KB	2660 MHz	5,2 h	1,2 h	5 GB	10302,234	104.66 art./sec
is	31,1	117	33009420 KB	2660 MHz	44,8 min	6,41 min	1 GB	778,189	104.66 art./sec

Table 3.2: Main cost parameters of history parsing

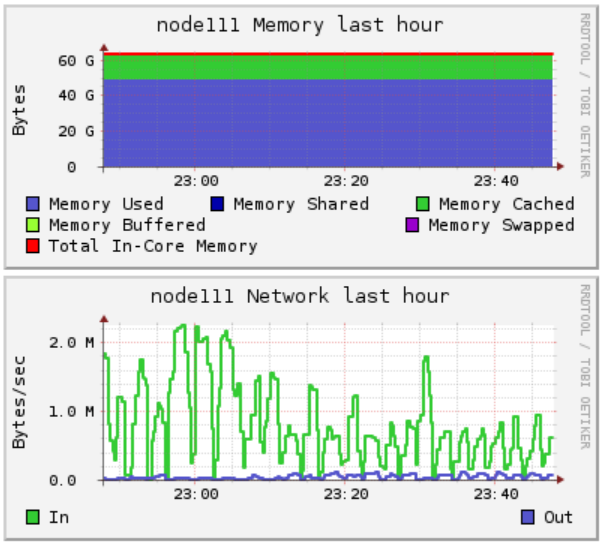

Figure 3.4: Memory consume for the german history parser process

On Table 3.3 we can see how PageRank process consumes a big amount of RAM memory due to the loading of links. However, this is a process with a CPU high performance and the amount of time it is used is higher in relation to the wall clock time probably because of the optimization of the process manager.

In Figure 3.5 for japanese PageRank process there are three different graphics with the CPU, memory and network use. First there is a period on which the node runs the process accumulating data and processing it with high peaks and later it gets to a stable point from which it will not change until it stores one value per article.

PageRank	Node	RAM	CPU	time wclock	time CPU	estimated RAM	mem GB·CPU
de	111	66103024 KB	2926 MHz	2,84 h	2,41 h	28 GB	11693,477
ja	112	33019548 KB	2992 MHz	2,51 h	2,1 h	16 GB	25831,1
ca	117	33009420 KB	2660 MHz	0,66 h	0,38 h	8 GB	4066,571
is	117	33009420 KB	2660 MHz	4 min	1 min	0,5 GB	174,06

Table 3.3: Main cost parameters of PageRank process

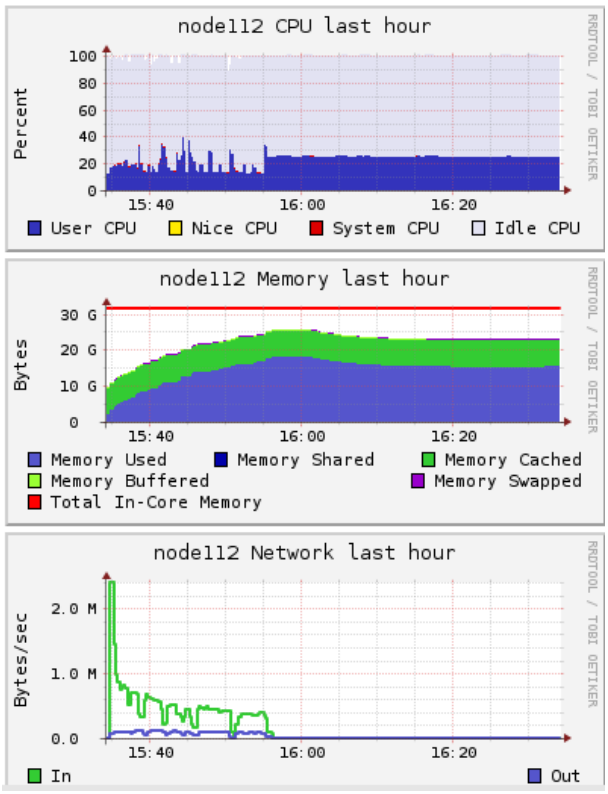

Figure 3.5: CPU, Memory and Network consume for Japanese PageRank process

All in all, the main restriction - memory RAM - makes not possible to process big repositories of languages like English or German in any of the cases, the last articles or the history file. Medium languages like Japanese and also small like Catalan it is possible to parse and process them without the history features. And small ones like Icelandic can be parsed in both kinds and processed for all characteristics - including PageRank - with relatively few amount of time.

Therefore the scope of the tool is reduced to research scenarios where there is resources enough. Although we must remark at the same time that research in small languages is not very developed in the current literature and also that they represent around 240 out of the 275 languages in which Wikipedia has editions.

CHAPTER 4. RESULTS

In this chapter there are the obtained results for the evaluation of Autoreferentiality. The initial section called Hyperlingua will review the indicators in the twenty selected languages according to the previously set hypothesis. Those which are confirmed and can reflect the the difference raised by the selected set of articles will be correlated to see their uniqueness. Eventually, they will be merged an index.

In section Viquipèdia the reader may find different listings from the most representative articles in the catalan edition selection. These will be produced by changing their weights and ranking them according to their importance in each dimension and thus visualizing the actual core of the set. Some other indicators and data analysis not treated in this chapter have been left in the Appendix A.

4.1. Hyperlingua

The semantic dimension refers to the scope of the selection of articles. Thus it is the most important and it will imply different results on the indicator analysis. In table 4.1 there are the number of articles for the whole language edition, all the selected levels and the effective levels (0-3) on which the indicators will be studied.

The average value of the effective levels is 24,89%, being the Japanese with largest extension in percentage with 52,06% and the Indonesian the least, with 12,19%. In absolute values a language like Guarani which is the second in percentage becomes the smallest. Since Autoreferentiality is measured as a language edition property the coverage is the parameter.

Lang.	Nº Art. Eff. Lvl	Nº Art. Sel. Lvl.	Nº Art. L.Edit.	% Sel. To L.Edit.	% Eff. To Sel.	% Eff. To L.Edit.	Pos.
ar	31424	51356	134253	38,25	61,19	23,41	13
ca	42413	51585	301304	17,12	82,22	14,08	19
cs	47266	50485	184251	27,40	93,62	25,65	8
da	43947	133194	141767	93,95	32,99	31,00	4
fi	55724	56600	261678	21,63	98,45	21,29	14
gn	526	526	1371	38,37	100,00	38,37	2
he	31755	34139	114496	29,82	93,02	27,73	6
hu	42928	45992	182467	25,21	93,34	23,53	12
id	18228	18960	149509	12,68	96,14	12,19	20
is	10434	10455	42023	24,88	99,80	24,83	10
it	115329	141489	767906	18,19	81,51	14,83	17
ja	383721	427233	737085	57,96	89,82	52,06	1
ko	40904	42672	155256	27,48	95,86	26,35	7
nl	96413	104496	650733	16,06	92,26	14,82	18
no	57061	77997	290629	26,84	73,16	19,63	15
ro	48303	48740	155763	31,29	99,10	31,01	3
sv	107232	115597	382801	30,20	92,76	28,01	5
sw	5061	5084	21193	23,99	99,55	23,88	11
tr	30362	44943	155242	28,95	67,56	19,56	16
zh	85441	123671	334175	37,01	69,09	25,57	9

Table 4.1: Number of articles by sets (effective levels, all levels and language edition)

The acronyms mean: Nº Art. Eff. Lvl. (*number of articles obtained through the first four levels 0-3 of selection*), Nº Art. Sel. Lvl. (*number of articles obtained by all the levels of selection*), Nº Art L.Edit. (*Total number of articles in the language edition*), % Sel. To L. Edit (*Percentage in extension of the number of articles from the selected levels to the language edition*), % Eff. To Sel. (*Percentage in extension of the number of articles from the effective levels on the number of articles from the selected levels*), % Eff. To L.Edit. (*Percentage in extension of the articles from the effective levels on the total articles from the language edition*), Pos. (*Position*).

It can be concluded that there is a will on representing local content. The difference on content between language editions can be on main topics but yet the articles which present the own culture, activities or heritage are represented in a not inconsiderable amount. Also, it is universal. It takes a similar piece of the encyclopedia besides the sociological context of the language, its size in Wikipedia and community activity.

4.1.1. Hypothesis testing

To test all the hypothesis we will follow the presented methodology. Each indicator value is obtained through the differece between its average on the set and on the whole language edition set of articles in relation to this last average. An indicator will be confirmed when the average of all the indicator values for each language is positive.

It is important to remark this procedure of relating to the all language edition articles result since it is the general trend. Thus it is what gives meaning to a minor change on the set. Besides, since local content is gathered through a selection using the graph the influence may be different in each level. These results will be developed on Section A.2 Level Analysis in Appendix A.

4.1.1.1. H1. Isolation

The first dimension is measured by the feature interwiki links, which show if an article is replicated in other languages. Table 4.2 shows how the hypothesis is confirmed in all cases. The selected articles are clearly much more isolated since they present an average of five times less links outside to other language editions.

The language edition presenting a lower average of interwiki links is the Japanese, both in the set of articles and all the language edition articles. Although the biggest difference is in the Finnish whose articles generally have around eight interwiki links but only one if it is on local content.

A very small language like Guarani has the highest number of interwiki links in both cases. This can be a consequence of the creating articles on the basis of another language article. Further details on interwiki links can be found in Appendix A section A.2 Isolation extra indicators where there is an analysis on the direction of the interwiki links.

Languages	Avg. Set	Avg. L.Edition	Diff.	Val. ind.	Pos.
ar	3,1	7,7	4,6	59,8	18
ca	1,4	6,4	5,0	78,6	11
cs	1,7	8,3	6,6	79,1	10
da	2,5	9,0	6,5	71,8	14
fi	1,0	8,0	7,0	87,4	1
gn	10,7	16,9	6,2	36,7	20
he	3,0	10,1	7,1	70,2	16
hu	2,8	8,0	5,2	65,4	17
id	0,9	7,1	6,2	87,0	2
is	1,3	8,8	7,4	84,7	4
it	2,5	4,9	2,4	49,5	19
ja	0,7	3,7	3,0	80,0	9
ko	1,2	8,1	6,9	85,4	3
nl	1,2	5,5	4,3	78,4	12
no	1,0	6,3	5,3	84,2	5
ro	1,4	7,9	6,5	82,6	6
sv	1,2	6,4	5,2	81,7	7
sw	2,9	14,6	11,7	80,2	8
tr	2,2	7,5	5,3	70,7	15
zh	1,4	5,8	4,4	75,7	13

Table 4.2: Results for the indicator interwiki links

The acronyms mean: Avg. Set (*Average of the feature from the selected set of articles*), Avg. L.Edition (*Average of the feature on the language edition*), Diff. (*Difference*), Val. Ind. (*Value of the indicator*), Pos. (Position).

The average of all the indicator values is a group value or a weight which shows the meaningfulness of an indicator. Later this value because it is positive it can be used for the index creation.

Weighting for the indicator interwiki links
74,4

4.1.1.2. H2. Effort

Effort is measured by the two indicators bytes and outlinks. According to Table 4.3 we can see how the first of them shows a great variability across languages in set of articles and all articles. In only eight of twenty the local content is larger in average, which means it is not possible to conclude that the indicator is good and a longer article reflects a higher importance by its editors.

Despite it does not show the expected difference, further analysis on levels may show more information which may make clear if level zero (including the key words in title) does accomplish the hypothesis at least. Instead, it can be said that the positive average of indicators as Weighting make it possible to include it in the index.

Languages	Avg. Set	Avg. L.Edition	Diff.	Val. ind.	Pos.
ar	3452,0	3102,2	349,7	11,3	6
ca	3368,0	4130,6	-762,7	-18,5	19
cs	4196,2	4595,6	-399,4	-8,7	17
da	2920,7	3225,6	-304,9	-9,5	18
fi	3631,7	3768,7	-137,0	-3,6	12
gn	3976,9	3443,5	533,4	15,5	5
he	5087,3	4676,1	411,1	8,8	7
hu	5309,3	5577,2	-267,9	-4,8	14
id	4031,6	3194,2	837,4	26,2	3
is	3181,5	2349,8	831,7	35,4	2
it	6899,4	4450,4	2.449,0	55,0	1
ja	3268,2	3323,4	-55,2	-1,7	10
ko	2672,8	2617,6	55,2	2,1	8
nl	4212,4	3372,3	840,1	24,9	4
no	2625,0	2867,8	-242,8	-8,5	16
ro	3230,4	3332,2	-101,8	-3,1	11
sv	2769,7	2880,7	-111,0	-3,9	13
sw	2212,1	2946,4	-734,3	-24,9	20
tr	3946,0	4001,2	-55,2	-1,4	9
zh	2598,0	2759,4	-161,4	-5,8	15

Table 4.3: Results for the indicator bytes

Weighting for the indicator bytes
4,2

The following indicator outlinks obtains a higher value for each language and its hypothesis is accomplished for eighteen cases (Table 4.4). The positive result in the weighting rises to 21,2, which is not as large as in the interwiki links indicator but much larger than bytes. The greater differences in the set and all articles is reflected on languages like Italian, Guarani or Indonesian.

It is clear then that those articles referring local content contain more links towards other articles. Editors make them more complete by linking them to other content. Also, there is a possibility on correlation between both effort indicators since the longest the article the more chances there are outlinks.

Languages	Avg. Set	Avg. L.Edition	Diff.	Val. ind.	Pos.
ar	13,7	11,2	2,5	22,3	8
ca	23,5	27,7	-4,1	-15,0	20
cs	29,9	31,1	-1,3	-4,1	19
da	22,5	20,2	2,3	11,2	12
fi	22,6	21,4	1,2	5,4	16
gn	9,0	5,3	3,7	69,3	2
he	49,9	39,6	10,3	26,0	7
hu	35,7	31,9	3,8	12,0	11
id	24,2	15,9	8,4	52,7	3
is	11,7	10,5	1,1	10,9	13
it	51,5	30,4	21,2	69,7	1
ja	51,6	44,2	7,3	16,6	9
ko	33,4	23,3	10,1	43,6	4
nl	29,0	21,3	7,7	36,3	5
no	21,1	19,7	1,3	6,7	15
ro	15,6	15,6	0,0	0,3	18
sv	23,3	21,2	2,1	9,8	14
sw	11	10,9	0,1	1,3	17
tr	24,4	21,0	3,4	16,0	10
zh	38,5	28,8	9,7	33,7	6

Table 4.4: Results for the indicator outlinks

Weighting for the indicator outlinks
21,2

4.1.1.3. H3. Prominence

Prominence dimensions take into account relational characteristic of Wikipedia. Features like inlinks, PageRank and category memberships can be used to look at the prominence from the articles in the whole language edition (External prominence) or within the same selected set (Internal prominence or endogamy). Thus we are analyzing at the same time the inlinks indicator for both two dimensions/kinds of prominence.

In Table A.4 we can check how the articles from the set are not more prominent than the average and thus the hypothesis is not confirmed. Only in cases like Indonesian or Guarani present a higher number of inlinks. Instead when used to detect their endogamy and see how many of them come from the same set all of the languages are over the fifty percent.

Languages like Swahili or Japanese are very endogamyc and their local content is defined by their own terms with few contact with other topics from the encyclopedia. However, this indicator was expected to be higher than fifty percent as it is the minimum for considering endogamy. Thus the difference between this value and 50 is the rellevant information and what is compared to 50 to get the Indicator Value.

Lang.	Avg. Set Inl.	Avg. LEdit.	Diff.	Val. Ind.	Pos.	Avg. Set End.Inl.	Perc.	Diff.	Val. Ind.	Pos.
ar	13,56	11,17	2,38	21,3	4	8,21143	60,6	10,6	21,2	19
ca	15,59	27,69	-12,10	-43,7	20	10,5493	67,7	17,7	35,3	15
cs	20,60	31,15	-10,55	-33,9	19	15,64	75,9	25,9	51,9	8
da	15,62	20,24	-4,62	-22,8	14	11,55	74,0	24,0	47,9	11
fi	16,49	21,44	-4,95	-23,1	15	12,61	76,5	26,5	53,0	7
gn	7,17	5,34	1,83	34,3	2	5,44	75,9	25,9	51,8	9
he	32,45	39,55	-7,10	-18,0	11	25,00	77,1	27,1	54,1	6
hu	21,50	31,89	-10,39	-32,6	17	15,04	70,0	20,0	40,0	14
id	24,23	15,88	8,34	52,5	1	13,44	55,5	5,5	11,0	20
is	8,18	10,60	-2,42	-22,8	13	6,13	75,0	25,0	50,0	10
it	37,65	30,54	7,11	23,3	3	23,66	62,9	12,9	25,7	16
ja	40,07	44,25	-4,18	-9,5	8	34,15	85,2	35,2	70,5	2
ko	21,94	23,26	-1,32	-5,7	7	18,05	82,3	32,3	64,6	3
nl	21,79	21,28	0,52	2,4	6	15,58	71,5	21,5	43,0	12
no	13,24	19,81	-6,57	-33,2	18	8,22	62,1	12,1	24,2	17
ro	11,45	15,57	-4,12	-26,5	16	9,41	82,3	32,3	64,5	4
sv	16,52	21,20	-4,68	-22,1	12	12,92	78,2	28,2	56,4	5
sw	9,01	10,88	-1,87	-17,2	10	7,93	88,1	38,1	76,1	1
tr	18,30	21,00	-2,71	-12,9	9	11,31	61,8	11,8	23,7	18
zh	30,47	28,82	1,65	5,7	5	21,47	70,5	20,5	40,9	13

Table 4.5: Result for the indicators inlinks and inlinks from set (endogamy)

Weighting for the indicators inlinks	Weighting for the indicators inlinks from set
-9,2	49

Endogamy of inlinks has a certain correlation with isolation. The more endogamy of inlinks the less the content is related with the rest of encyclopedia and also with other language editions. In Figure 4.1 there is a graphical representation of the most important concepts for the catalan selection of local content. These are the most inlinked and therefore those which work better to define other articles.

Figure 4.1: Cloud with the most inlinked articles from the same set (endogamy)

Using the feature category memberhips we discover it is both confirmed as an external prominence indicator and endogamy. Table 4.6 shows us how in all cases the average from the set is higher than the one from all articles. Editors want their content to be identified with more labels to be more precise.

Also, the same indicator when analysed the percentage of category memberships from the set articles obtained through the selected categories shows a similar value than the other endogamy indicator. It can be concluded that this content is labeled mostly as local content. However both endogamy indicators are not correlated.

Lang.	Avg. Set CM.	Avg. LEdit.	Diff.	Val. Ind.	Pos.	Avg. Set End.CM.	Perc.	Diff.	Val. Ind.	Pos.
ar	2,55	2,13	0,4	19,6	18	1,67	65,83	15,8	31,7	17
ca	2,35	1,54	0,8	52,3	6	1,91	81,54	31,5	63,1	4
cs	3,46	2,71	0,7	27,5	14	2,23	64,68	14,7	29,4	18
da	2,71	1,98	0,7	36,5	12	2,57	95,02	45,0	90,1	2
fi	2,39	2,11	0,3	13,0	20	1,64	68,91	18,9	37,8	15
gn	1,04	0,91	0,1	14,3	19	0,98	95,41	45,4	90,8	1
he	3,33	2,31	1,0	43,9	9	2,69	80,91	30,9	61,8	5
hu	3,05	2,13	0,9	43,3	11	2,13	70,01	20,0	40,0	13
id	2,09	1,03	1,1	103,6	1	1,60	76,91	26,9	53,8	9
is	1,45	0,90	0,6	61,3	4	1,32	91,21	41,2	82,4	3
it	2,06	1,20	0,9	72,5	2	1,62	78,90	28,9	57,8	7
ja	3,81	3,17	0,6	20,4	16	2,69	70,59	20,6	41,2	10
ko	3,73	2,48	1,3	50,4	7	2,49	66,98	17,0	34,0	16
nl	2,32	1,62	0,7	43,6	10	1,80	77,66	27,7	55,3	8
no	3,50	2,38	1,1	47,1	8	1,95	55,80	5,8	11,6	19
ro	2,39	1,78	0,6	33,7	13	1,67	70,36	20,4	40,7	11
sv	3,52	2,78	0,7	26,7	15	1,91	54,32	4,3	8,7	20
sw	1,98	1,65	0,3	20,4	17	1,37	69,49	19,5	39,0	14
tr	3,29	1,93	1,4	70,2	3	2,30	70,03	20,0	40,1	12
zh	2,95	1,91	1,0	54,2	5	2,36	80,15	30,2	60,3	6

Table 4.6: Results for the indicators category memberships and category memberships from set (endogamy)

Weighting for the indicator category memberships	Weighting for the indicator category membership from set (endogamy)
42,7	53,2

Last, the indicator PageRank is calculated through the inlinks as it is explained in previous sections. In Table 4.7 we can see how it is not confirmed in half of the cases. Besides some exceptions like Swedish and Dutch, the set articles present a lower average on PageRank value.

Lang.	Avg. Set	Avg. L.Edit.	Diff.	Indicator Value	Pos.
ar	5,73E-06	7,36E-06	-1,62E-06	-22,1	15
ca	2,49E-06	3,38E-06	-8,91E-07	-26,3	17
cs	4,82E-06	5,40E-06	-5,75E-07	-10,7	12
da	5,68E-06	7,01E-06	-1,33E-06	-19,0	13
fi	3,98E-06	3,80E-06	1,71E-07	4,5	9
gn	6,75E-04	6,34E-04	4,13E-05	6,5	8
he	6,84E-06	8,68E-06	-1,83E-06	-21,1	14
hu	7,19E-06	5,46E-06	1,73E-06	31,7	5
id	1,02E-05	6,54E-06	3,68E-06	56,3	4
is	2,11E-05	2,27E-05	-1,54E-06	-6,8	11
it	1,24E-06	1,22E-06	2,59E-08	2,1	10
ja	2,29E-06	1,35E-06	9,40E-07	69,7	3
ko	4,93E-06	6,39E-06	-1,46E-06	-22,8	16
nl	2,83E-06	1,53E-06	1,30E-06	85,5	2
no	4,04E-06	3,12E-06	9,16E-07	29,3	6
ro	3,79E-06	6,30E-06	-2,51E-06	-39,9	18
sv	5,39E-06	2,59E-06	2,80E-06	108,3	1
sw	1,65E-05	4,67E-05	-3,02E-05	-64,7	20
tr	3,49E-06	6,28E-06	-2,79E-06	-44,4	19
zh	3,57E-06	2,96E-06	6,09E-07	20,6	7

Table 4.7: Results for the indicator PageRank value

Weighting for the indicator PageRank value
6,8

4.1.1.4. H4. Edition

Edition dimension is similar to effort but it measures the actions and actors, edits and editors, instead of length and outlinks. The first indicator number of edits is not confirmed for twelve cases. In Table 4.8 we can appreciate so much variability among languages, where Icelandic turns out to have many more edits in their local content than the average of all articles and hebrew while it is the one with more articles in average for all articles. The average value or weighting is 0,7.

Lang.	Avg. Set	Avg. L.Edit.	Diff.	Indicator Value	Pos.
ar	17,20	25,72	-8,5	-33,1	14
ca	18,90	16,17	2,7	16,9	7
cs	17,85	23,81	-6,0	-25,0	11
da	21,78	23,89	-2,1	-8,9	8
fi	15,78	27,37	-11,6	-42,4	16
gn	16,47	26,54	-10,1	-37,9	15
he	29,08	51,15	-22,1	-43,1	17
hu	13,76	31,72	-18,0	-56,6	19
id	21,03	17,17	3,9	22,5	6
is	43,16	16,53	26,6	161,2	1
it	58,33	30,57	27,8	90,8	3
ja	15,85	39,04	-23,2	-59,4	20
ko	19,73	26,49	-6,8	-25,5	12
nl	11,19	25,32	-14,1	-55,8	18
no	15,39	19,41	-4,0	-20,7	10
ro	16,29	20,26	-4,0	-19,6	9
sv	17,75	24,97	-7,2	-28,9	13
sw	39,33	18,66	20,7	110,7	2
tr	50,96	35,75	15,2	42,6	4
zh	38,30	30,12	8,2	27,1	5

Table 4.8: Results for the indicator edit count

Weighting for the indicator edit count
0,7

Second indicator, number of editors, is not confirmed in all cases either. Eleven out of the twenty cases give a positive indicator value and their order in ranking previsibly is similar to number of edits. The firsts show a great distance in relation to the following. The average value or weithing is 30,3 which is a good indicator for the index.

Lang.	Avg. Set	Avg. L.Edit.	Diff.	Indicator Value	Pos.
ar	13,38	12,69	0,7	5,5	10
ca	8,60	8,48	0,1	1,3	11
cs	14,62	13,67	1,0	7,0	9
da	12,67	13,98	-1,3	-9,4	13
fi	13,15	15,44	-2,3	-14,9	14
gn	8,50	11,90	-3,4	-28,5	17
he	18,44	24,62	-6,2	-25,1	16
hu	14,01	14,31	-0,3	-2,1	12
id	15,57	9,39	6,2	65,9	3
is	30,01	7,99	22,0	275,7	2
it	25,59	15,59	10,0	64,2	4
ja	11,98	22,10	-10,1	-45,8	20
ko	15,09	12,26	2,8	23,1	7
nl	9,55	14,78	-5,2	-35,4	19
no	13,15	10,56	2,6	24,4	6
ro	7,61	10,99	-3,4	-30,8	18
sv	12,55	14,77	-2,2	-15,0	15
sw	33,63	8,62	25,0	289,9	1
tr	25,80	18,79	7,0	37,3	5
zh	16,27	13,65	2,6	19,2	8

Table 4.9: Results for the indicator editor count

Weighting for the indicator editor count
30,3

Last indicator on edition is diversity coefficient to represent the relations between number of edits and number of editors. Yet is clear that the higher value of them means more interest but it is important too to know how the content is created. We expected a lower diversity in edition (few editors more active) and it is confirmed by sixteen out of twenty.

Language editions like Hungarian or Swahili show how a majority of eighty percent of their local content is created by just 41% and 44% of the editors who contributed. A language like Icelandic shows how its

Wikipedia is made from few but their local content is much more consensed in diversity of edits. In average for all languages the weighting is 8,6.

Lang.	Avg. Set	Avg. L.Edit.	Diff.	Indicator Value	Pos.
ar	0,72	0,81	0,09	11,0	7
ca	0,72	0,79	0,07	9,0	8
cs	0,76	0,81	0,04	5,4	12
da	0,80	0,81	0,00	0,5	15
fi	0,72	0,79	0,07	8,9	9
gn	0,80	0,72	-0,08	-11,1	19
he	0,74	0,71	-0,03	-4,7	17
hu	0,41	0,71	0,30	42,0	2
id	0,80	0,73	-0,07	-9,9	18
is	0,70	0,59	-0,11	-19,0	20
it	0,70	0,75	0,04	5,8	11
ja	0,68	0,79	0,11	14,1	5
ko	0,75	0,75	0,00	0,0	16
nl	0,57	0,81	0,23	29,0	3
no	0,68	0,74	0,06	8,2	10
ro	0,68	0,83	0,15	18,4	4
sv	0,71	0,81	0,10	11,9	6
sw	0,44	0,81	0,37	45,8	1
tr	0,76	0,78	0,02	2,4	14
zh	0,71	0,74	0,03	3,7	13

Table 4.10: Results for the indicator diversity coefficient

Weighting for the indicator diversity coefficient
8,6

Further analysis on edition dimension can be found in Appendix A section A.3 with Edition extra indicators. There is an analysis on edits and editors by type to understand how the community is composed and how they are related to local content. Also, the diversity coefficient concept is applied to the whole set of articles and the language edition.

4.1.1.5. H5. Temporal

Temporal dimension wants to measure how the interest on this local content can be measured along time compared to the whole language edition. First, in Figure 4.2 we can see how the Catalan language edition grows in number of articles each semester in both all articles and selected ones. In general, the set follows the language edition trend in all languages.

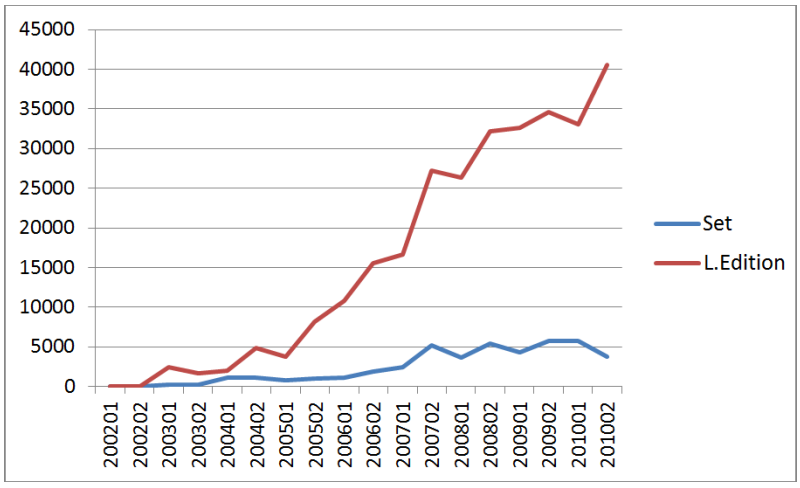

Figure 4.2: Temporal growth in semesters of Catalan language edition since its creation

To understand if there is a higher relative growth we measure two rates of creation related to the set and to all articles. The comparison of these rates gives place to first indicator, percentage of times the rate is superior, and the comparison of the increments between periods of the rate gives place to the second indicator, incremental rate. This can indicate a change in trend.

In Table 4.11 there are the percentage of times in which the set rates and increments are superior to the whole language edition. According to our condition to measure the indicator which is the set should have larger rates more than fifty percent of the times (this is: higher activity than language edition) they are not confirmed in general but few cases.

In the superior relative indicator we can see values around forty percent of the cases which means that the set stays in the general trend but does not imply higher motivation in creation. In the incremental we can see how in languages like Hungarian the set rate is not going down as much as the whole language edition rate or it advances to the general trend. It is 80% of the times higher.

Both indicators give a negative result in average and thus they cannot be used in the index creation. However the temporal dimension can be analysed in greater detail in the Appendix A section A.4 on different treatments on rates and levels indicating interest fluctuation and antiquiy based predictor.

Lang.	% Rate Su.	% Incr. Rate. Su.	Diff. Rate.	Diff. Incr.	Val. Ind. Rate	Po.	Val. Ind. Incr.	Po.
ar	37,69	33,85	-12,31	-16,15	-24,62	16	-32,31	19
ca	40,68	44,92	-9,32	-5,08	-18,64	11	-10,17	3
cs	43,85	33,08	-6,15	-16,92	-12,31	3	-33,85	9
da	42,31	24,62	-7,69	-25,38	-15,38	6	-50,77	18
fi	43,85	25,38	-6,15	-24,62	-12,31	4	-49,23	16
gn	29,23	17,69	-20,77	-32,31	-41,54	20	-64,62	1
he	37,69	30,00	-12,31	-20,00	-24,62	17	-40,00	12
hu	41,54	80,00	-8,46	30,00	-16,92	8	60,00	2
id	39,23	42,31	-10,77	-7,69	-21,54	14	-15,38	20
is	40,77	30,77	-9,23	-19,23	-18,46	10	-38,46	10
it	42,37	56,78	-7,63	6,78	-15,25	5	13,56	15
ja	41,54	20,77	-8,46	-29,23	-16,92	9	-58,46	6
ko	42,31	35,38	-7,69	-14,62	-15,38	7	-29,23	11
nl	40,00	13,85	-10,00	-36,15	-20,00	12	-72,31	17
no	40,00	26,92	-10,00	-23,08	-20,00	13	-46,15	7
ro	34,62	16,15	-15,38	-33,85	-30,77	19	-67,69	5
sv	44,62	30,00	-5,38	-20,00	-10,77	2	-40,00	4
sw	38,46	29,23	-11,54	-20,77	-23,08	15	-41,54	14
tr	36,15	43,85	-13,85	-6,15	-27,69	18	-12,31	13
zh	45,38	81,54	-4,62	31,54	-9,23	1	63,08	8

Table 4.11: Results for the indicators of creation rates

Weigthing for the indicator rate dif.	Weighting for the indicator rate increment
-19,77	-28,29

4.1.2. Indicator evaluation

In Figure 4.3 there is a table with a summary for the indicators of the six dimensions. The highest values are for Interwiki links from Isolation (74.45), Inlinks from set (45,30) and Categories Memberships from set (48,48) from Endogamy, and Editor count from Edition (30,33). Inlinks and both Temporal indicators are negative and then discarded for the index.

The three other rows show different variations on the indicators results. The first is calculated by means of standard deviation in Indicator Values for all languages, the second is the value from the standard deviation of Avg. Set divided by the average of Avg. Set. The third is the same but on the Avg. L.Edition. The first works to know universality of indicators comparing them and the second and third to know where it variates more.

The first shows how interwiki links and diversity coefficient are the indicators which present less variation across languages and thus can be considered the most universals in Autoreferentiality. However, Interwiki

links when looking at the variation in Set Average and All Articles Average has more variation in the set than in all articles.

	Isolation	Effort		External Prom.		
Indicator	Interwiki	Outlinks	Bytes	Inlinks	Cat. Memberships	PageRank
Weigthing	74,45	21,24	4,25	-9,21	42,73	6,84
Variation (Ind.V.)	14,05	20,79	17,68	22,29	23,37	39,73
Variation Set Art.	1,28	0,46	0,35	0,42	0,28	29,58
Variation All Art.	0,41	0,42	0,27	0,42	0,30	24,52

Internal Prom.		Edition		80%	Temporal	
Endo. Inlinks.	Endo. C.M.	Edit count	Editor count	Div. Coeff.	Rel. Rate	Incr. Rel. Rate
45,30	48,48	0,74	30,33	8,58	-19,77	-28,29
22,59	26,53	52,48	80,77	14,49	30,15	45,80
0,53	0,24	0,72	0,51	0,14		
		0,32	0,31	0,07		

Figure 4.3: Summary table with all the values from the evaluated indicators

The one which shows more variation in the indicator values is Editor count. Then there is edit count, which supports the idea that edition is very diverse depending on the language and community. All in all, the variation is always bigger in the selected set of articles than in all articles from language editions.

Having stated this about the indicator the last checking before the index is made by correlation statistics. In Table 4.12 4.13 we can see all the indicators correlated among themselves (included edit and editor count by type). Therefore, after checking it in languages editions like Italian, Czech and Romanian we can see usual correlations between indicators.

da(s)	IW	B	OL	IL	CM	PR	ILFS	CMFS	EC	ETC	DC	ECU	ECIP	ECB	ETCU	ETCIP	ETCB
IW	1,00	0,23	0,13	0,18	0,02	0,12	0,14	0,03	0,33	0,34	-0,06	0,20	0,19	0,48	0,24	0,17	0,43
B	0,23	1,00	0,74	0,24	0,21	0,07	0,24	0,21	0,34	0,25	-0,11	0,35	0,26	0,17	0,26	0,20	0,15
OL	0,13	0,74	1,00	0,25	0,28	0,07	0,26	0,27	0,32	0,23	-0,12	0,34	0,24	0,14	0,26	0,19	0,11
IL	0,18	0,24	0,25	1,00	0,03	0,17	0,98	0,03	0,28	0,22	-0,04	0,26	0,20	0,20	0,21	0,18	0,13
CM	0,02	0,21	0,28	0,03	1,00	0,02	0,03	0,98	0,10	0,07	-0,10	0,11	0,06	0,05	0,10	0,05	0,02
PR	0,12	0,07	0,07	0,17	0,02	1,00	0,15	0,02	0,25	0,17	-0,04	0,17	0,22	0,26	0,13	0,14	0,13
ILFS	0,14	0,24	0,26	0,98	0,03	0,15	1,00	0,03	0,27	0,21	-0,03	0,27	0,21	0,17	0,21	0,18	0,11
CMFS	0,03	0,21	0,27	0,03	0,98	0,02	0,03	1,00	0,10	0,08	-0,10	0,11	0,06	0,05	0,10	0,05	0,03
EC	0,33	0,34	0,32	0,28	0,10	0,25	0,27	0,10	1,00	0,64	-0,22	0,93	0,83	0,62	0,61	0,58	0,36
ETC	0,34	0,25	0,23	0,22	0,07	0,17	0,21	0,08	0,64	1,00	-0,19	0,54	0,58	0,46	0,90	0,86	0,67
DC	-0,06	-0,11	-0,12	-0,04	-0,10	-0,04	-0,03	-0,10	-0,22	-0,19	1,00	-0,23	-0,14	-0,12	-0,19	-0,15	-0,13
ECU	0,20	0,35	0,34	0,26	0,11	0,17	0,27	0,11	0,93	0,54	-0,23	1,00	0,73	0,36	0,60	0,52	0,18
ECIP	0,19	0,26	0,24	0,20	0,06	0,22	0,21	0,06	0,83	0,58	-0,14	0,73	1,00	0,30	0,56	0,67	0,17
ECB	0,48	0,17	0,14	0,20	0,05	0,26	0,17	0,05	0,62	0,46	-0,12	0,36	0,30	1,00	0,27	0,22	0,66
ECTU	0,24	0,26	0,26	0,21	0,10	0,13	0,21	0,10	0,61	0,90	-0,19	0,60	0,56	0,27	1,00	0,78	0,40
ETCIP	0,17	0,20	0,19	0,18	0,05	0,14	0,18	0,05	0,58	0,86	-0,15	0,52	0,67	0,22	0,78	1,00	0,28
ETCB	0,43	0,15	0,11	0,13	0,02	0,13	0,11	0,03	0,36	0,67	-0,13	0,18	0,17	0,66	0,40	0,28	1,00

Table 4.12: All the indicators averages correlation for the selected articles in Danish language edition

The acronyms mean: IW (*interwiki links*), OL (*outlinks*), B (*bytes*), IL (*inlinks*), CM (category memberships), PR (PageRank), EIL (Endogamy *inlinks*), ECM (Endogamy category memberships), EC (edit count), ETC (editor count), DC (diversity coefficient), ECU (edit count users), ECIP (edit count ip users), ECB (edit count bots), ECTU (editor count users), ETCIP (editor count ip), ETCB (editor count bots).

da(e)	IW	B	OL	IL	CM	PR	ILFS	CMFS	EC	ECT	DC	ECU	ECIP	ECB	ETCU	ETCIP	ETCB
IW	1,00	0,19	0,13	0,15	-0,06	0,13			0,48	0,57	-0,18	0,21	0,17	0,72	0,29	0,21	0,81
B	0,19	1,00	0,73	0,21	0,14	0,13			0,45	0,38	-0,18	0,50	0,34	0,25	0,40	0,34	0,24
OL	0,13	0,73	1,00	0,37	0,22	0,19			0,45	0,38	-0,18	0,51	0,34	0,23	0,42	0,34	0,21
IL	0,15	0,21	0,37	1,00	0,02	0,55			0,47	0,42	-0,09	0,44	0,36	0,33	0,41	0,41	0,25
CM	-0,06	0,14	0,22	0,02	1,00	0,01			0,10	0,09	-0,11	0,16	0,10	0,00	0,17	0,10	-0,01
PR	0,13	0,13	0,19	0,55	0,01	1,00			0,33	0,31	-0,07	0,29	0,24	0,27	0,29	0,28	0,21
ILFS																	
CMFS																	
EC	0,48	0,45	0,45	0,47	0,10	0,33			1,00	0,95	-0,31	0,88	0,80	0,75	0,86	0,83	0,70
ETC	0,57	0,38	0,38	0,42	0,09	0,31			0,95	1,00	-0,26	0,77	0,75	0,79	0,89	0,82	0,79
DC	-0,18	-0,18	-0,18	-0,09	-0,11	-0,07			-0,31	-0,26	1,00	-0,28	-0,18	-0,28	-0,22	-0,17	-0,25
ECU	0,21	0,50	0,51	0,44	0,16	0,29			0,88	0,77	-0,28	1,00	0,76	0,40	0,85	0,80	0,36
ECIP	0,17	0,34	0,34	0,36	0,10	0,24			0,80	0,75	-0,18	0,76	1,00	0,33	0,74	0,94	0,29
ECB	0,72	0,25	0,23	0,33	0,00	0,27			0,75	0,79	-0,28	0,40	0,33	1,00	0,52	0,39	0,95
ETCU	0,29	0,40	0,42	0,41	0,17	0,29			0,86	0,89	-0,22	0,85	0,74	0,52	1,00	0,82	0,49
ETCIP	0,21	0,34	0,34	0,41	0,10	0,28			0,83	0,82	-0,17	0,80	0,94	0,39	0,82	1,00	0,35
ETCB	0,81	0,24	0,21	0,25	-0,01	0,21			0,70	0,79	-0,25	0,36	0,29	0,95	0,49	0,35	1,00

Table 4.13: All the indicators averages correlation for all the articles in Danish language edition

We can see that PageRank does not have a good correlation with any other indicator but those related to Edit count. This makes sense since the PageRank as it is explained before can be considered as the probability of stepping into an article and therefore more possibilities of editing it.

Diversity coefficient is very uncorrelated with the others and the least is edit count. It can be pointed the more edits the most probable there is diversity. Bytes have a slight correlation with edit and editor count. The more activity the longer the articles.

Finally, we can identify four clear different couples in correlation: bytes-outlinks, inlinks-inlinks from set, category memberships-category memberships from set and edit count-editor count. Then in selecting the definitive indicators for the index we choose between them and the criteria is the one which took higher value as indicator.

The list of selected indicators is: interwiki links (Isolation), bytes (Effort), category memberships (Prominence), inlinks from set (Endogamy), editors and diversity coefficient (Edition).

4.1.3. Autoreferenciality index

The index is the sum of the selected indicators, their values and selected articles coverage for each language. The calculation is explained in section 2.2.1 after the analysis dimension but briefly it consist in weighting the individual values multiplying by a general weight across languages and the extension of the set in the language edition.

In Figure 4.4 there is the table with all the resultant values and final result. The language edition with highest autoreferentiality is the Icelandic (with a high diversity in editing but many editors too) and the least is the Catalan. The second one is the Japanese which stands out for being very isolated and endogamyc. Both with very different combination in the addition of indicators.

INDEX	Isolation	Effort	External Prom.	Internal Prom.	Edition	80%
Indicator	Interwiki	Outlinks	Category Mem.	Endo. Inlinks.	Núm. Editors	Div. Coeff.
Weighting	74,45	21,24	42,73	45,30	30,33	8,58

Languages							Results	Position
is	15,65	0,57	6,50	5,63	20,77	-0,41	**48,71**	1
ja	31,01	1,83	4,54	16,63	-7,23	0,63	**47,41**	2
sw	14,26	0,07	2,08	8,24	21,00	0,94	**46,58**	3
ko	16,75	2,44	5,68	7,71	1,85	0,00	**34,43**	4
ro	19,06	0,02	4,46	9,07	-2,89	0,49	**30,21**	5
da	16,57	0,74	4,84	6,73	-0,88	0,01	**28,01**	6
sv	17,04	0,58	3,19	7,16	-1,28	0,29	**26,98**	7
he	14,50	1,53	5,21	6,80	-2,11	-0,11	**25,82**	8
cs	15,11	-0,22	3,02	6,03	0,55	0,12	**24,60**	9
gn	10,49	5,65	2,34	9,00	-3,32	-0,37	**23,80**	10
hu	11,45	0,60	4,36	4,26	-0,15	0,85	**21,36**	11
tr	10,29	0,67	5,87	2,10	2,21	0,04	**21,17**	12
no	12,30	0,28	3,95	2,15	1,46	0,14	**20,27**	13
fi	13,85	0,25	1,18	5,12	-0,96	0,16	**19,60**	14
id	7,89	1,36	5,40	0,60	2,44	-0,10	**17,59**	15
it	5,47	2,19	4,59	1,73	2,89	0,07	**16,94**	16
ar	10,41	1,11	1,96	2,24	0,39	0,22	**16,33**	17
zh	2,19	1,83	5,92	4,74	1,49	0,08	**16,26**	18
nl	8,64	1,14	2,76	2,89	-1,59	0,37	**14,21**	19
ca	8,24	-0,45	3,14	2,25	0,06	0,11	**13,35**	20

Figure 4.4: Table with all the values and final index

4.2. Viquipèdia

In this last section we consider useful to show different rankings of articles ponderated according to the different dimesions of Autoreferentiality in catalan language edition. This way the reader may understand better which kind of articles are those analyzed in the study and how they fit different characteristics.

Differently than the index construction, each indicator value (average of the whole language edition) weights the articles multiplying by their values. The sum of all the values may give a result to rank. Last, varying the weightings with different values choosen on purpose will give emphasis to one or other dimension creating a typology.

4.2.1. Top 20 most rated articles

First result is obtained through the same language indicators weighting. Those which were negative or could not reach 10% have been turned to this value. The two endogamic indicators have been normalized with the inlinks and category memberships. Date, which is a time indicator will not be taken into account.

From the following list one can observe public figures and associations with a relative social rellevance. These are articles which have obtained high values in isolation (not many interwiki links) and are mostly endogamy. This means that they have no interest out of the catalan sphere and they can be critized from the notability point of view.

weighting	74.6	10	10	10	52.3	10	35.33	63.08	16.8	10	10	0
Title	IW	OL	B	IL	CM	PR	EIL	EAC	EC	ECT	CD	D
Joan Ainaud de Lasarte	1	65	13525	38	15	2.45E-6	38	14	150	32	0,355	30/11/2005
Josep Maria Ainaud de Lasarte	0	104	17336	32	13	2.13E-6	29	13	127	31	0,433	31/07/2007
Rafael Blasco Castany	0	90	8989	41	9	2,82E-02	41	9	67	14	0,385	15/09/2008
Ricard Bellveser Icardo	1	50	4753	5	9	1.39E-6	5	9	4	1	0	19/07/2010
Unió Deportiva Torredembarra	0	25	3898	8	16	7.50E-7	7	12	7	1	0	14/09/2009
Grup Excursionista i Esportiu Gironí	1	60	15316	51	15	3,43E-02	49	12	73	22	0,381	13/06/2004
Agustí Altisent Altisent	0	42	3908	9	9	1.26E-6	8	9	81	17	0,25	29/04/2007
Aigua d'Ora	0	92	16108	68	8	9.13E-6	66	8	67	7	0,333	29/08/2008
Bonaventura Gassol i Rovira	1	195	11580	69	12	9.2E-6	60	12	54	30	0,69	28/04/2007
Eugenio Burriel de Orueta	0	33	4045	9	9	9,04E-04	9	9	34	14	0,615	21/02/2009
Joan Francesc Mira i Casterà	1	43	5569	56	17	5.80E-6	33	16	85	40	0,589	28/08/2005
Jordi d'Ornós	0	54	3196	5	9	9.72E-7	5	9	20	11	0,699	14/09/2007
Riera de l'Hospital	0	43	7697	12	6	1.88E-6	12	6	7	1	0	19/10/2010
Cèsar Panicot i Llagostera	0	17	2574	1	6	5.46E-7	1	6	9	1	0	04/03/2010
Verònica Cantó Domènech	0	17	2538	1	6	9.36E-7	1	6	2	1	0	10/08/2010
Inma Sancho	0	6	1872	2	6	6.77E-7	2	6	4	1	0	28/03/2010
Massís de Bonastre	0	47	3048	1	9	5.90E-7	1	9	12	7	0,833	06/03/2009
Francesc Almela i Vives	0	91	6866	13	9	1.40E-6	12	9	22	9	0,625	25/12/2007
Jordi Balló i Fantova	0	27	5144	3	7	1.19E-6	3	7	54	10	0,444	10/07/2007
Daniel Giralt-Miracle i Rodríguez	0	41	3705	11	8	1,67E-02	11	8	33	18	0,706	16/07/2007
Carles Dénia Moreno	0	57	5062	2	5	5.49E-7	2	5	3	1	0	21/07/2010

Table 4.14: Top 10 articles with more score in the set

The acronyms mean: IW (*interwiki links*), OL (*outlinks*), B (*bytes*), IL (*inlinks*), CM (category memberships), PR (PageRank), EIL (Endogamy *inlinks*), ECM (Endogamy category memberships), EC (edit count), ETC (editor count), D (data).

4.2.2. Typology

Using the previous weighting method we present a typology of a ranks for each dimension. Each indicator obtains a value of ten unless it belongs to the selected dimensions to put emphasis, which wIW receive the equivalent to the sum of the other indicators values. Hence we assure emphasis on one dimension.

At the table 4.15 one can observe the selected articles according to the interwiki links. They are also large (Bytes), very inlinked and have a low coefficient of diversity.

weighting	110	10	10	10	10	10	10	10	10	10	10	10
Title	IW	OL	B	IL	CM	PR	EIL	EAC	EC	ETC	CD	D
Llista de plantes de Catalunya	0	3892	202958	7	3	4.99E-6	3	2	98	16	0,267	30/09/2006
Història del ferrocarril a Catalunya	0	722	56799	28	2	1.80E-5	19	2	465	42	0,098	12/12/2005
Història del Dret català	0	535	26852	94	1	3,58E-01	86	1	333	34	0,061	03/12/2005
Llista de jaciments arqueològics de Cat.	0	2071	100201	76	4	9.21E-6	72	3	353	7	0,167	28/09/2010
Llista d'escuts del País Valencià	0	904	64488	2	1	1.91E-6	2	1	220	31	0,167	03/11/2008
Josep Maria Ainaud de Lasarte	0	104	17336	32	13	2.136E-6	29	13	127	31	0,433	31/07/2007
Falugues	0	1	627	1	1	3.45E-6	1	1	1	1	0,000	30/09/2005
Joan Ainaud de Lasarte	1	65	13525	38	15	2.44E-6	38	14	150	32	0,355	30/11/2005
Llista d'espais naturals del País Val.	1	688	30630	50	2	4.50E-6	50	2	260	33	0,125	13/09/2005
Manuel Brunet i Solà	0	423	40789	13	4	1.33E-6	11	4	241	14	0,077	12/09/2008
Llista de Creus de Sant Jordi	0	1301	48824	845	2	9.19E-5	841	1	265	54	0,283	04/12/2006

Table 4.15: Top 10 articles with more score according to external interest

The following type of articles have been selected according to internal interest (bytes and outlinks). At the table 4.16 one can see articles which contain different entities defined in other articles (lists). In some cases a low diversity in edition is also clear.

weighting	10	100	100	10	10	10	10	10	10	10	10	10
Title	IW	OL	B	IL	CM	PR	EIL	EAC	EC	ETC	CD	D
Llista de plantes de Catalunya	0	3892	202958	7	3	4.9E-6	3	2	98	16	0,27	30/09/2006
Llista de plantes del País Valencià	0	2965	277000	6	2	5,7E-02	1	1	13	10	0,89	06/04/2008
Llista de plantes de les IWes Balears	0	1892	168462	5	3	5,6E-02	0	2	43	7	0,5	07/04/2008
Senyera Reial	3	890	175235	152	3	1.15E-5	100	2	1185	92	0,01	01/09/2005
Llista de noms de plantes de Catalunya	1	2275	109401	2	3	5.38E-7	1	2	105	19	0,28	26/04/2007
Futbol Club Barcelona	26	1440	127067	4384	2	1.78E-4	2646	1	1806	295	0,25	28/09/2003
Llista de jaciments arqueològics de Cat.	0	2071	100201	76	4	9.21E-6	72	3	353	7	0,17	28/09/2010
Corona d'Aragó	9	999	162163	1515	2	2.90E-4	788	1	788	171	0,29	10/09/2004
Gran Companyia Catalana	8	981	147497	189	1	7.40E-6	125	1	1068	40	0,05	08/10/2006
Llista de municipis de Catalunya	4	1893	74607	24	1	6.44E-5	21	1	96	50	0,63	02/04/2004

Table 4.16: Top 10 articles with more score according to internal interest (bytes/outlinks)

At the table 4.17 there are those which have been selected according to inlinks, PageRank and endogamy of inlinks. They are rellevant articles according to the catalan context. Barcelona, Catalunya o València they are very important and also well rated according to other dimensions. Hence it is no strange they are also replicated in other repositories.

weighting	10	10	10	90	10	90	90	10	10	10	10	10
Titol	IW	OL	B	IL	CM	PR	EIL	EAC	EC	ETC	CD	D
Barcelona	26	901	98785	19484	2	0,0023	12772	1	1413	519	0,458	29/10/2003
Catalunya	26	600	69926	9721	1	0,00195	6453	1	673	203	0,391	06/11/2003
Generalitat de Catalunya	13	160	17387	2661	1	5.91E-4	2178	1	238	94	0,505	27/09/2003
València	26	445	50029	5067	1	6.22E-4	3170	1	752	235	0,410	17/07/2003
País Valencià	25	597	55436	4397	2	6.55E-4	2524	1	1088	246	0,282	15/04/2003
Llista de Creus de Sant Jordi	0	1301	48824	845	2	9.19E-5	841	1	265	54	0,283	04/12/2006
Xarxa ferroviària de Catalunya	2	711	49276	549	1	4.82E-5	544	1	286	29	0,071	07/08/2008
Palmarès del Futbol Club Barcelona	2	1004	48365	2	1	7.38E-7	2	1	442	74	0,192	15/08/2006
Llista d'espais naturals del País Val.	1	688	30630	50	2	4.50E-6	50	2	260	33	0,125	13/09/2005
Joan Ainaud de Lasarte	1	65	13525	38	15	2.44E-6	38	14	150	32	0,355	30/11/2005

Table 4.17: Top 10 articles with more score according to Prominence-Endogamy of inlinks and PageRank

Most rated articles concerning category memberships (endogamic too) are again public figures 4.18. They are usually categorized according to biographic aspects, social groups, etc. Sometimes they are also well rated in number of inlinks or diversity coefficient.

weighting	10	10	10	10	100	10	10	100	10	10	10	10
Title	IW	OL	B	IL	CM	PR	EIL	EAC	EC	ETC	CD	D
Joan Francesc Mira i Casterà	1	43	5569	56	17	5.80E-6	33	16	85	40	0,590	28/08/2005
Joan Ainaud de Lasarte	1	65	13525	38	15	2.44E-6	38	14	150	32	0,355	30/11/2005
Josep Maria Ainaud de Lasarte	0	104	17336	32	13	2.13E-6	29	13	127	31	0,433	31/07/2007
Carles Riba i Bracons	5	222	15452	152	15	2.83E-5	91	13	160	73	0,569	19/12/2003
Josep Maria Espinàs i Massip	2	117	11178	56	12	4.50E-6	43	12	85	37	0,556	28/04/2005
Lluís Companys i Jover	15	195	20147	284	12	7.15E-5	216	12	279	120	0,546	04/01/2004
Grup Excursionista i Esportiu Gironí	1	60	15316	51	15	3,43E-02	49	12	73	22	0,381	13/06/2004
Bonaventura Gassol i Rovira	1	195	11580	69	12	9.19E-6	60	12	54	30	0,690	28/04/2007
Josep Benet i Morell	1	81	6151	29	12	9.53E-6	22	12	50	22	0,619	17/06/2007
Joan Fuster i Ortells	5	138	21823	182	10	4.37E-5	142	10	298	99	0,418	09/08/2004

Table 4.18: Top 10 articles with more score according to Prominence-Endogamy in category memberships

Most edited articles are confirmed as those which raise much interest as well. In table **??** we can see how those with more edits and editors, low edition diversity are also those more prominent in links. Some repeated titles are Barcelona and País Valencià.

Ponderació	10	10	10	10	10	10	10	10	90	90	90	10
Títol	IW	OL	B	IL	CM	PR	EIL	EAC	NE	NET	CD	D
Barcelona	26	901	98785	19484	2	0,00231823	12772	1	1413	519	0,458	29/10/2003
Futbol Club Barcelona	26	1440	127067	4384	2	1.78041E-4	2646	1	1806	295	0,252	28/09/2003
Senyera Reial	3	890	175235	152	3	1.15915E-5	100	2	1185	92	0,011	01/09/2005
País Valencià	25	597	55436	4397	2	6.55118E-4	2524	1	1088	246	0,282	15/04/2003
Jaume el Conqueridor	19	936	99681	1481	13	2,96E+00	786	9	923	203	0,272	04/03/2003
Gran Companyia Catalana	8	981	147497	189	1	7.40451E-6	125	1	1068	40	0,051	08/10/2006
Club Esportiu Atlètic Balears	8	198	59673	60	2	2.40981E-6	45	2	1026	44	0,023	16/02/2004
Mercè Rodoreda i Gurguí	10	695	107067	127	5	3,40E-01	82	4	865	152	0,166	23/04/2004
Metro de Barcelona	19	685	74728	588	2	4.75973E-5	551	1	963	225	0,308	30/08/2004
Sabadell	18	384	40777	1481	1	1.64407E-4	1278	1	830	225	0,371	20/03/2004

Table 4.19: Top 10 articles with more score according to edition (edit and editor count and diversity coefficient)

Last, those which are rated as the oldest can be understood as articles which have interest among time. Since many articles are coincident in a similar period of creation there are some which are already appeared before. It is curious to see in table A.18 article Granollers is previous to the article Catalunya.

Ponderació	10	10	10	10	10	10	10	10	10	10	10	110
Títol	IW	OL	B	IL	CM	PR	EIL	EAC	EC	ETC	CD	D
Barcelona	26	901	98785	19484	2	0,00231823	12772	1	1413	519	0,458	29/10/2003
Catalunya	26	600	69926	9721	1	0,00190255	6453	1	673	203	0,391	06/11/2003
Futbol Club Barcelona	26	1440	127067	4384	2	1.78041E-4	2646	1	1806	295	0,252	28/09/2003
Jaume el Conqueridor	19	936	99681	1481	13	2,96E+00	786	9	923	203	0,272	04/03/2003
Catedral de Girona	5	454	64238	161	6	2.58868E-5	130	5	474	75	0,162	29/09/2003
Granollers	15	51	11921	604	1	5.54727E-5	405	1	193	101	0,630	12/04/2002
Gran Teatre del Liceu	11	928	89313	779	1	5.91582E-5	505	1	375	80	0,266	09/08/2003
Girona	23	671	58680	1996	1	2.88474E-4	1594	1	685	245	0,443	26/08/2003
València	26	445	50029	5067	1	6.2295E-4	3170	1	752	235	0,410	17/07/2003
Reus	16	464	45973	1103	1	1.35664E-4	935	1	544	214	0,498	28/06/2003

Table 4.20: Top 10 articles with more score according to oldest in creation

CHAPTER 5. CONCLUSIONS

In this last chapter called conclusions there will be two sections summarizing the main points of the document and setting the future lines. For this it will be necessary to discuss both the development according to the goals and the use case particular conceptualization and results.

Since the nature of the object is both technical and social it will be necessary to consider the importance of its impact. On the community and also in the development of applications which are based in collaborative knowledge repositories. Finally, there will be proposals for extensions and new tests in which the study may continue.

5.1. Discussion and achieved goals

5.1.1. Technical

How can Wikipedia be analyzed as a technical object with an information architecture? This is a key question initial to this research and with great interest from social sciences studies. In fact, the deep understanding of the structures, its meaning and weight, could be used in models to determine the behaviour or governance of the massive project. Wikipedia was in turn analysed by several software, and the same supporting Wikimedia Foundation incited its research, but there were not enough integral approaches accounting all its characteristics. Neither a good bridging from the results to the community.

In this context, the development of the existing tool wikAPIdia aimed giving an answer to these lacks and obtained positive results. Likewise, it also gave account of the requirements raised by the large information treatment. wikAPIdia was revised from the theory to be continued and extended for a broader use, covering more scenarios and setting mechanisms for automatization. The technical characterization from the repository was the same important as reviewing the existing technical disciplines which used the data for their research. For they recognise the information as quantitative, textual and relational. Then, the most important characteristic was Wikipedia language edition compatibility of the tool, which was also extended until cover thirty.

In this sense, the tool becomes much more versatile than any other in a general-purpose mature technology like Java and MySQL. It includes new indispensable features. wikAPIdia new version can analyse the history of Wikipedia with all the relations between edits, editors and their types. Also the restructuring of the information into the database and the procedures to obtain them was essential for it increased its optimization in avoiding repetition. Summary features now include what is most important from the repository and display it related to the main element the article.

Also, previous literature showed certain aspects in the creation of the repository like a very active critical mass or the importance of few articles to define the rest. This suggested us to include new features like a coefficient to measure the diversity of edition in an article as well as in a set of them, or also, the PageRank calculation. wikAPIdia provides methods for enable a programmer treating most important elements from the repository and analyze it to extract their features.

We realized after this development that on another version of wikAPIdia we might have to decide whether including alternative spaces or developing new features semantically related from Natural Language Processing area. Besides, these new features would require storing text, which is something wikAPIdia avoids by locating the piece in the XML in a search when it is necessary as it was not its main purpose. Neither it is to do semantic tagging according to their languages like for instance DBpedia project do with considerable success.

Then, once the tool can cover up to most of the problematics from Wikipedia we may look back at the technical characterization and see if what we did not implement can be interesting for future improvements. Spaces like ”Discussion” are not interesting for the casual readers and represent a particular need for the editors. For instance, they would be useful for a very focused study on controversy which would complement edition changes with discussion analysis. Portals, Wikiprojects and other information structures which classify articles by their topic are similar cases. It might be better to include features which give more information on the

temporal analysis. These could be creation rates (peak, average, effective, etc.) in order to create complex models for growth prediction.

In short, wikAPIdia is a tool which permits information retrieval and cultural data mining. For that, we must consider it has been essential to have complete use case which allowed us to try all the tool features and functionalities and identify real problems in the experimentation. Because Autoreferentiality as a cultural-related concept has been tried in all sizes of Wikipedia language editions. This scale testing permitted debugging in the implementation as well as required a long-term often use of the tool.

When extending the case use to larger data files you may find problems like exceeding RAM memory heap, database indexing or just interminable iterations which may be approached by other means. And all the derived problems from crash risk or process cost, examined in detail. However, the solutions we proposed for security used replication by different means and did not present any consistency problem. Instead, when increasing the number of Wikipedia language editions and tests to apply to them, the human control and attention needed raises. This other problem was solved by a well structured abstraction, an argument code for tests and a script definition for remote class calls.

However, considering wikAPIdia complete did not stop us from developing classes to the data externalization in order to continue the analysis or presentation with other tools. This was important for all the non-java programmed software which gives more functionalities like in our case Matlab or Excel. For this, methods to extract series of features and their multiple possibility normalizations have been developed. Also database tables for the results to be consulted from a webpage. Notwithstanding, these have not been used with the same regularity as others.

5.1.2. Scientific

Use case ”Cultural Configuration” was choosen after being involved in Catalan Wikipedia in a survey for the motivation. Among the possible answers it presented, the one which obtained more votes was something like ”I like writing about this topic”, beyond ideological and other reasons common to the motivation studies. However, many editors gave as a second reason for writing ”to make Catalan culture visible in the Internet and enriching the language”. The national topic was also present in questions about conflict or interest. Despite, it was never treated by any literature and the goal of an encyclopedia was gathering knowledge with any priority.

Usually, motivation was approached by classic social sciences methodologies which discuss about where it resides, in the individual or in it while acts. Further than that, an analysis on the content cannot provide a clear answer on motivation but it can explain if the interests from the editors are confirmed. Cultural configuration stands out to explain the diversity from the same project in different languages and how belonging to a nation or language eventually pushes part of the editors to write about their local content. While most of the research assume the results obtained from English language extrapolable to other language editions, this study remarks how the difference exist and it is important to those who create the product.

We divided the problem into the scope of the local content and its characteristics with the goal of comparing languages. Autoreferentiality, as a concept which expresses the degree of interest, has been broken down into different dimensions. Eventually, This divide and conquer approach answers two interesting dependent questions, what is local and does it matter, and allows examining at its result independently. The first showed us that a good method for selecting articles was needed and the more unbiased one was using the same process by which the editors create the content and classify them. From general (key words of the main territory, gentilic and language) to the particular.

This method had the advantage of using the same Wikipedia structure and thus understanding each level gathered as a particularization but still related to the core of the national or language meaning. The heterogeneity in topics which could be related to the words was found, besides some interference which were neutralized by selecting just the first three levels after the keywords. Results explained these articles where in average a thirty percent of the repositories. Other methods which could reduce the interference were evaluated, like rating all the articles according to the relevance of few words, but not implemented. This is proposed to be a future line of study.

All in all, this key word selection method was confirmed by natives from each language and the margin of interpretation is almost null. Other studies were like Self-focus bias was based on geographical articles which were common across language editions. By increasing the selection we are making the study semantically related and we can understand the interest the information suggest in other languages as an indicator. Effort, prominence, endogamy, edition and temporal, describe the editors from all Wikipedia features perspectives and thus allow seeing or not the difference in interest.

Dimensions included indicators which were conceptually showing similar aspects. Then it has been no surprise to realize sometimes they are correlated (and include just one in the index). For the results also showed that some indicators were more stable across languages than others, or in other words local content despite being about different topics and different cultures has universal features. More features related to the new spaces like discussion or digging into the history from articles to create new coefficients would be indicators on the same dimensions. In this case, they could be effort and edition.

It is pertinent to ask about the usefulness of merging the indicators into an index. However, after considering the coommonalities and the relations between indicators, this could explain more about the languages we measured. Using the weightings on the indicators was in favor of the behaviour of the indicator in all language editions and therefore not to introduce any bias. Once the indicators are measured, other compositions of the index can be obtained. Applying the same idea of an index into the articles of Catalan language edition was merely illustrative but helped in understanding the local content composition.

Other important analysis was an in-depth levels of selection using the same Autoreferentiality indicators. Surprisingly, the hypothesis of higher interest proved all right even those not confirmed for the selected set of articles. Also, looking at the indicators evolve through the levels proved the interest is semantically related. They have different value depending on the distance from the core level with the keywords. Along the levels there were some oscillations in the values repeated in all languages, which permits us to say that there are transition levels and more specific into the topic.

Indicators like editor count and diversity coefficient showed an interesting point in how the length of articles is related to the definition of the article scope. The variation along levels in length was high, being the first ones (general articles with keywords) longer. But the editor count showed not a big oscillation (standard deviation), which means that the same amount of editors were doing a long article or a short one depending on the degree of specialization. Diversity coefficient showed also that the first levels had a subgroup of editors much more active which was doing most of the editing. From this it can be concluded that for making a language edition grow you need to promote the specialization in articles, for the general articles they will be large and complete anyway.

Also on Autoreferentiality we observed the edits and editors by their type (user, ip or bot) and the temporal stability. A great majority of the local content was made by a thirty percent of the whole registered users community. And one every third anonymouse decided to do it in this kind of articles too. This is rellevant since all the motivation studies have consulted registered users and this shows there is no need for a commitment or involving in the community to have this local content related motivation. The analysis confirmed which role played each of them in the creation of Wikipedia, but leaves for future work topics like a deep understanding of bots, like for instance when their activity starts in an article.

The extra analysis on temporal data can be summarized with two indicators, the stability of the interest on local content as well as the antiquity. The first was analyzed by using their creation rate and statistics. The second showed how the articles which included keywords in their title or were closer to them were much older, which mean that the local content from a language edition is not likely to grow much more. The temporal analysis of Wikipedia is something which has not be researched by the academic community and possibly there will appear some literature in the near future.

5.1.3. Social

Having a social impact was not a goal in itself, but considering the characteristics of the study object and use case the results needed to be explained to the community. Wikipedia is done by everyone but the community consens its rules, define the categories and point the direction in which the encyclopedia goes. Publishing a survey, topical coverage or this local content study bring awareness to the individuals which are usually

focused on their articles or tasks within the project.

This work explains something which may seem obvious but only after we could only affirm after seeing results. All the cultures when they explain human knowledge tend to dedicate bigger effort to explain themselves and where they are located in it. Catalan language edition was interested in the study since the beginning by means of its association Amical Viquipèdia and proposed Wikimedia Chapter ”Wikimedia CAT”. After the survey which explicitly asked on ”national motivation”, the results prove Catalan edition contrary to that has not a higher Autoreferentiality degree than others but the lowest one.

To explain the conclusions of this study and its use case we were invited to the tenth anniversary cycle of events. First, on a round table on Wikipedia and research, remarking the possibilities of wikAPIdia and the relational and quantitative analysis. And later to the community in a international meeting with other language editions representatives. Some of the results, concepts and points discussed were expanded to a greater detail in order to bring empirical evidence for the new promotional plans among those attending language editions. Specially those regarding edition and temporal dimensions. Languages like Esperanto, Occitan, Basque or Aragonese were processed to obtain the results for this event.

However, the current broad hyperlingual approach has defined in a way of looking forward to the presentation at Wikimania, which will take place in Haifa, August 3rd to 5th. All the tests can be expanded with more statistical analysis in order to make it more complete and sound to the world Wikipedia community event. Meanwhile this same document can be found in the wiki (`http://www.lsi.upc.edu/~mmiquel/mediawiki/index.php/Main_Page`). We are open to improvement and suggestions to extend it to more languages and tests.

5.2. Conclusions and future lines

This document has presented wikAPIdia, a Wikipedia analytical software, and the use case about cultural influence on the repositories, ”Autoreferentiality”. The java tool is built as an extension from a previous version to include several new functionalities, from new features and language compatibility to ready environment for processing. We have structured the software to cope with the requirements from a technical characterization of Wikipedia. Also the design had to deal with problems derived from large data sets (XML and later MySQL) and to automatize mechanisms for multiple testing.

Along with the different chapters we have seen how important was for the design to analyze what was valuable for the research in Wikipedia. Not all the spaces and information were interesting and a way of abstracting their relations is an important point for approaching diverse problematics. In addition, the comprehension of the technical fields in which Wikipedia is analyzed (NLP, DM and IR) made us include features and prepare the tool to extract data into different syntaxis, like for instance Weka Data Mining tool. Also, specific tables in the database to retrieve results from a webpage or a wiki.

The tool was used in all the developed functionalities and with multiple objects in the social problematic previously sketched ”Autoreferentiality” measurement. Its goal of understanding the degree of interest and extension of the local content was pertinent and interesting in the Wikipedia context, but also in the current Internet where the user has become the main information producer. The work also fitted an empty gap from the current motivation literature which influenced by other environments like free software never took into account in Wikipedia ”national” or self-representation as reasons for writing.

The study proved the local content represents in average a quarter of the encyclopedias, regardless the size in number of articles, the number of speakers, the world location or cultural root. Small languages without an expanded use did not show higher extension in local content, like it could be a cultural heritage preservation. Yet, it has been shown which community is more active, consistent and the way they develope the articles. The selected articles having in common a national meaning were more categorized and unique in the sense of non replicated. Web applications relying on Wikipedia (the biggest collaborative resource) cannot take it as one same project in different languages and may know what is in it in order to take profit of it.

Also, the trend in putting higher interest in local content reflected on the categorization process Wikipedia is interesting considering the Semantic Web, in which the editor needs to tag for the machines to understand and display the content differently, more contextually. Also, Autoreferentiality has shown how few concepts

in each language are important to all the local content (territories, cities, associations, political parties) and how the degree of endogamy in defining the local content among themselves was slightly correlated to the isolation of this content from other repositories.

Wikipedia in English is the leader in number of articles, but also the one which allocates more local content from other language editions. However, it only represents a small part counting all the Wikipedia language editions projects. Therefore it is possible to study the difference in the same content from many points of view, since the goals are the same but not the perspectives. Wikipedia provides this chance and wikAPIdia is a tool successfully designed for this purpose allowing treatment and analysis without starting from scratch.

Some of the future lines which we have pointed to in the discussion will be published as papers. Another selection of local content by different methodologies will aim at proving the similar extension it may take but overall understanding of its thematic composition and relation to the culture key concepts. While this document is written there is another one in process about the higher categorization of local content and its influence on Semantic Web whose conclusions were already presented in Dresden, 21-26th February, conference on ”Cross Modal Analysis of Verbal and Nonverbal Communication”.

However, the biggest challenge will be presenting some of these conclusions regarding Autoreferentiality and the whole editing community to the event Wikimania with all the language editions communities. After having introduced them to Catalan Wikipedia community in special events and to other geographically near languages, there has been a good response for using the points as feedback for promotion. This has been a project which became bigger as we realized of the nature of its object, its technical requirements when approaching it and scientifical needs when defining a problematic, for it will continue developing and in both three ways.

ACKNOWLEDGEMENTS

The work presented in this research would not have been possible without the help and encouragement transmitted by a great number of friends, colleagues, family and professors. Most of them are listed in the following lines:

Eduard Aibar, Joan Campàs, Horacio Rodríguez, Sebastià Sallent, Marcos Faúndez and Joan Gomà. Gabriel Verdejo and Iván Couto. Ilina Kareva, Kazuhiro Tsubono, Ma Li, Marco Grassi and Koen Van Doorslaer.

Diana Petri. David Morera and Pere Tuset. Jordi Montfort.

Fina Ribé, Jordi Miquel, Joan Ribé, Emília Díez, Agustí Miquel, Dolors Cotonat and Peius Cotonat.

TERMINOLOGY AND ACRONYMS

Technical - Wikipedia acronyms and abbreviations

API Application Programming Interface. 2, 11, 29, 30, 31, 55.
CPU Central Process Unit. 36, 37, 38, 39.
DM Data Mining. 10, 35.
GNU Recursive acronym ”GNU’s Not Unix!”. 7.
ILL Interlanguage Links (see IW). 7.
IR Information Retrieval. 10.
IW Interwiki. 50, 51, 52, 53, 54.
NLP Natural Language Processing. 10, 11, 17.
PHD Philosophy Doctor. 3, 10, 11, 29.
PR PageRank. 50, 52, 53, 54.
RAM Random-Access Memory. 11, 36, 37, 38.
SFR Self-focus Ratio. 14.
TF-IDF Term Frequency-Inverse Document Frequency. 17, 18, 26.
TB Terabyte. 29, 33.
UTC Universal Time Coordinated. 6.
UWAG Universal Wikipedia Article Graph. 9.
WAG Wikipedia Article Graph. 9.
WCG Wikipedia Category Graph. 9.
WP Wikipedia. 3.
WT Wikipedia Text. 7, 10.
XML Extensible Markup Language. 1, 9, 10, 11, 15, 16, 30, 32, 37, 55, 58.

AR Arabic
CA Catalan
CS Czech
DA Danish
DE German
GN Guarani
FI Finnish
HE Hebrew
HU Hungarian
IS Icelandic
IT Italian
ID Indonesian
JA Japanese
KO Korean
NL Nederlands
NO Norwegian
RO Romanian
SV Swedish
SW Swahili
TR Turkish
ZH Chinese

Terminology (Wikipedia English language edition definitions)

Motivation: ”the driving force which help causes us to achieve goals. Motivation is said to be intrinsic or extrinsic.”
Operationalization: ”the process of defining a fuzzy concept so as to make the concept clearly distinguishable (in humanities) or measurable (in physicalist sciences) and to understand it in terms of empirical observations.”
Index: ”a statistical measure of changes in a representative group of individual data points.”
Graph: ”a collection of vertices or ’nodes’ and a collection of edges that connect pairs of vertices.”

BIBLIOGRAPHY

[1] Alexa. Top sites. http://www.alexa.com/topsites, January 2011.

[2] Timme Bisgaard Munk. Self-efficacy and self-esteem in a knowledge-political battle for an egalitarian epistemology in wikipedia. *Observatorio (OBS*)*, 3(4):13–34, 2009.

[3] Brian Butler, Elisabeth Joyce, and Jacqueline Pike. Don't look now, but we've created a bureaucracy: the nature and roles of policies and rules in wikipedia. In *CHI '08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 1101–1110, New York, NY, USA, 2008. ACM.

[4] Fundació Enciclopèdia Catalana. Fundació. http://www.enciclopedia-catalana.cat/ca/fundacio.htm, January 2011.

[5] Andrea Forte and Amy Bruckman. Why do people write for wikipedia? incentives to contribute to open-content publishing. group 05 workshop position paper. In *GROUP 05 Workshop: Sustaining Community: The Role and Design of Incentive Mechanisms in Online Systems. Sanibel Island, FL*, pages 6–9, 2005.

[6] Jim Giles. Internet encyclopaedias go head to head. *Nature*, 438(7070):900–901, December 2005.

[7] Alexander Halavais and Derek Lackaff. An analysis of topical coverage of wikipedia. *Journal of Computer-Mediated Communication*, 13(2):429–440, 2008.

[8] F. Maxwell Harper, Sherry Xin Li, Yan Chen, and Joseph A. Konstan. Social comparisons to motivate contributions to an online community. In *PERSUASIVE'07: Proceedings of the 2nd international conference on Persuasive technology*, pages 148–159, Berlin, Heidelberg, 2007. Springer-Verlag.

[9] Brent Hecht. wikapidia. http://collablab.northwestern.edu/wikapidia_api/, February 2011.

[10] Brent Hecht and Darren Gergle. Measuring self-focus bias in community-maintained knowledge repositories. In *C38;T '09: Proceedings of the fourth international conference on Communities and technologies*, pages 11–20, New York, NY, USA, 2009. ACM.

[11] Brent Hecht and Darren Gergle. *The tower of Babel meets web 2.0: user-generated content and its applications in a multilingual context*, page 291–300. ACM, 2010.

[12] JUNG. Java universal network graph framework, February 2011, HOWPUBLISHED =.

[13] Aniket Kittur, Ed H. Chi, and Bongwon Suh. What's in wikipedia?: mapping topics and conflict using socially annotated category structure. In *CHI '09: Proceedings of the 27th international conference on Human factors in computing systems*, pages 1509–1512, New York, NY, USA, 2009. ACM.

[14] Vivi Nastase and Michael Strube. Decoding wikipedia categories for knowledge acquisition. In *AAAI'08: Proceedings of the 23rd national conference on Artificial intelligence*, pages 1219–1224. AAAI Press, 2008.

[15] Felipe Ortega. *Wikipedia: A quantitative analysis*. PhD thesis.

[16] Naren B. Peddibhotla and Mani R. Subramani. Contributing to public document repositories: A critical mass theory perspective. *Organization Studies*, 28(3):327–346, March 2007.

[17] Ulrike Pfeil, Panayiotis Zaphiris, and Chee S. Ang. Cultural differences in collaborative authoring of wikipedia. *Journal of Computer-Mediated Communication*, 12(1), 2006.

[18] Cindy Royal and Deepina Kapila. What's on wikipedia, and what's not . . . ? *Soc. Sci. Comput. Rev.*, 27(1):138–148, 2009.

[19] Richard M. Ryan and Edward L. Deci. Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, 25(1):54 – 67, 2000.

[20] Jakob Voss. Measuring wikipedia. In *Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics*, Stockholm (Sweden), July 2005.

[21] Jimmy Wales. Wikimania. *Keynote talk*, Cambridge, Mass, 2006.

[22] Wikimedia. Founding principles. `http://meta.wikimedia.org`, January 2011.

[23] Wikimedia. Stats. `http://stats.wikimedia.org`, January 2011.

[24] Wikipedia. Wikipedia, wiki, statement of principles, five pillars, wikipedia guidelines, wikipedia policy, user access levels, list of wikipedias by sample of articles, list of wikipedias, information retrieval, data mining, natural language processing. `http://www.wikipedia.org`, January 2011.

[25] Heng-Li Yang and Cheng-Yu Lai. Motivations of wikipedia content contributors. *Comput. Hum. Behav.*, 26(6):1377–1383, 2010.

[26] Xiaoquan Zhang and Feng Zhu. Intrinsic motivation of open content contributors: The case of wikipedia. In *Workshop on Information Systems and Economics*, Evanston, IL, 2006.

APPENDIX

APPENDIX A. COMPLEMENTARY RESULTS

In this second appendix the reader may find complementary results. It is valuable to see how with similar indicators or applied differently, we may find new insights on the cultural configuration and Wikipedia language editions in general. Therefore, in first section ”Levels analysis” the hypothesis of interest divergence on the selected articles can be reexamined in detail for each level and dimension.

Other sections go further into the analysis of the already reviewed dimensions. This is the case of Isolation, Edition or Temporal extra indicators, which can explain much better the leadership of the language editions, the composition of the community, or the stability of motivation in the community.

A.1. Levels analysis

This section wants to explain the inner structure of the selected articles. Since they were gathered using a graph crawling methodology, which used the key words to define the zero level to start, the following ones are related to them but more specific. Then, when the distance from level zero is getting bigger at each level also the value from the indicators are expected to decrease. The higher interest is related to the local content semantic value.

In order to check the hypothesis again for each indicator we have selected the same twenty language editions. Therefore we can see if there is a trend among the levels and also if some of them are above the all articles average.

A.1.1. H1.

In Table A.1 we can see the evolution in the ten levels of the indicator interwiki links. The usual trend in the language is that level zero is higher than the next one. This happens because they contain the key words and therefore represent summary articles of the culture, which have more possibilities of being replicated in other language editions. After this it slowly increases for three or four levels and then decreases again.

However, we can appreciate for all languages how in almost any of the levels the average of interwiki links is higher than the average of all articles. When this happens is in articles at the last levels which represent a degree of specialization and distance from the semantic core with the keywords.

Lang.	0	1	2	3	4	5	6	7	8	9	10	Ef.	Sel.
ar	3,8	2,9	2,8	3,3	4,1	4,2	4,1	9,6	8,4	11,2	0,0	3,1	3,6
ca	1,4	1,3	1,3	1,5	2,5	1,9	2,5	3,0	9,7	0,0	0,0	1,4	1,7
cs	1,4	1,7	2,0	1,3	6,7	7,0	9,7	7,4	5,7	6,4	6,6	1,7	2,1
da	2,0	1,5	2,5	7,5	11,4	12,5	13,2	12,4	12,1	12,4	8,1	2,5	9,2
fi	1,2	0,9	1,0	2,2	2,5	5,7	0,9	0,6	0,0	0,0	0,0	1,0	1,0
gn	7,4	7,1	16,9	12,3	0,0	0,0	0,0	0,0	0,0	0,0	0,0	10,7	10,7
he	2,1	2,2	4,3	4,8	8,9	7,3	6,2	1,4	0,0	0,0	0,0	3,0	3,4
hu	2,5	1,7	5,2	7,1	5,4	3,7	0,7	0,6	0,0	0,0	0,0	2,8	2,9
id	1,2	1,0	0,7	0,8	1,3	2,0	0,0	0,0	0,0	0,0	0,0	0,9	1,0
is	1,3	1,4	1,4	0,5	0,2	0,0	0,0	0,0	0,0	0,0	0,0	1,3	1,3
it	2,3	1,8	2,5	3,1	4,8	7,1	7,6	6,9	5,0	4,0	3,3	2,5	3,1
ja	0,7	0,7	0,7	1,0	2,5	3,3	7,5	10,0	9,3	7,9	7,1	0,7	1,1
ko	1,4	1,1	0,9	3,0	3,0	6,7	5,6	6,9	4,5	5,1	0,0	1,2	1,3
nl	1,7	1,1	1,1	1,9	3,6	6,8	6,8	7,4	3,3	0,0	0,0	1,2	1,6
no	0,7	0,9	1,0	1,4	1,6	3,4	4,8	2,9	10,4	12,7	0,0	1,0	1,5
ro	2,2	1,4	1,4	0,9	12,9	5,1	15,6	0,0	0,0	0,0	0,0	1,4	1,5
sv	1,3	1,0	1,0	2,4	3,7	7,5	3,3	5,5	9,4	3,3	11,1	1,2	1,5
sw	6,0	4,6	4,6	0,2	19,0	3,0	0,0	0,0	0,0	0,0	0,0	2,9	3,0
tr	1,8	1,9	3,0	2,4	5,0	7,4	9,8	11,1	9,7	11,6	13,5	2,2	4,6
zh	1,2	1,4	1,5	1,4	1,6	3,0	3,5	4,2	3,9	2,9	3,5	1,4	1,9

Table A.1: Evolution for the indicator interwiki links at selection levels

A.1.2. H2.

In Table A.2 we can see the evolution in levels for indicator bytes. Comparing the values we find in level zero we can see how this time the indicator is confirmed for all languages. Not only it becomes true but doubles the average from all articles in all cases. This is because those articles containing the key words must are the ones which represent better the local content. At their following level their values decrease but still in most cases is higher than the average. After three more levels it start increasing again but not as high as the two first. This is the reason why in column selected the average is lower than the average from effective, which just comprises the four first levels (0-3).

Lang.	0	1	2	3	4	5
ar	4543,23	3764,89	3381,82	2765,44	2697,29	2281,56
ca	6826,57	3673,91	3142,97	2726,99	3770,12	3508,38
cs	8127,34	4718,11	3844,24	3215,26	4752,34	4027,82
da	5911,30	2882,99	2534,60	3128,48	3315,33	3638,61
fi	7509,00	3316,42	3772,65	3873,80	5608,91	5184,80
gn	6473,60	5340,67	2208,36	1183,57	0,00	0,00
he	8602,25	4840,60	4998,98	5022,78	4730,38	4180,37
hu	10576,52	4679,92	5795,68	5955,38	5508,09	4128,35
id	6812,19	3882,97	3696,57	3581,20	2880,84	3276,73
is	4299,43	3208,43	2799,27	1552,49	1431,62	0,00
it	9911,54	7851,91	6914,73	5515,44	6187,57	5307,00
ja	5863,89	3299,67	3145,13	3096,38	4523,90	4101,48
ko	4955,52	2640,69	2217,70	2550,24	2453,20	1963,54
nl	6958,91	3899,48	4161,07	5200,78	4340,48	4091,19
no	5874,94	2570,09	2434,03	2583,79	1844,44	3111,30
ro	7592,89	3464,88	2355,98	3593,45	5175,90	5636,53
sv	6035,01	2684,34	2699,23	2357,40	3088,18	2803,84
sw	4182,81	2395,18	2553,64	1636,63	5226,73	12031,00
tr	5462,53	3521,39	4243,36	4111,95	4294,53	3920,36
zh	4611,19	2740,64	2323,67	2315,13	2104,29	2211,47

Lang.	6	7	8	9	10	Effective	Selected
ar	2832,22	3588,37	3381,82	3094,12	0,00	3451,95	3107,07
ca	2642,06	2379,44	4235,89	0,00	0,00	3367,97	3390,31
cs	4328,29	4420,67	8980,57	2227,39	2206,13	4196,19	4210,90
da	3625,24	3505,98	3400,34	3808,87	3302,88	2920,73	3326,75
fi	2453,64	2455,42	0,00	0,00	0,00	3631,70	3655,57
gn	0,00	0,00	0,00	0,00	0,00	3976,85	3976,85
he	3680,76	5268,72	0,00	0,00	0,00	5087,27	5044,59
hu	2881,79	2811,11	0,00	0,00	0,00	5309,32	5313,67
id	0,00	0,00	0,00	0,00	0,00	4031,57	3989,90
is	0,00	0,00	0,00	0,00	0,00	3181,52	3178,00
it	3616,18	6209,78	4706,70	4868,07	3512,05	6899,39	6598,76
ja	2923,81	4220,01	4810,64	4603,88	3627,62	3268,23	3343,26
ko	2073,06	3265,20	1636,94	3171,63	0,00	2672,83	2657,98
nl	5005,51	5486,62	3094,27	0,00	0,00	4212,35	4234,09
no	3006,59	2347,37	6121,17	4125,78	0,00	2625,00	2575,58
ro	10744,44	0,00	0,00	0,00	0,00	3230,41	3251,08
sv	2464,46	2782,80	3321,58	3412,73	7190,43	2769,71	2777,74
sw	0,00	0,00	0,00	0,00	0,00	2212,11	2227,09
tr	5623,89	5369,76	4756,00	4665,82	4820,96	3945,97	4220,74
zh	3163,08	3628,60	3511,30	3050,44	3536,13	2598,00	2694,09

Table A.2: Evolution for the indicator bytes at selection levels

In Table A.3 we can see how the outlinks indicator responses similarly. The first levels almost double the average from all articles and are a fourty percent higher than the average from all levels. The same oscillation after few levels appears and last levels have high values, but not enough for the average of all levels to be higher than the average of the three. However this proves again certain correlation between bytes and outlinks.

Lang..	0	1	2	3	4	5	6	7	8	9	10	Ef.	Sel.
ar	18,51	14,41	13,58	11,29	10,79	9,13	11,74	17,76	17,06	13,88	0,00	13,67	12,41
ca	57,92	26,66	22,18	16,20	25,26	22,78	22,04	20,92	19,28	0,00	0,00	23,54	23,51
cs	74,03	34,78	26,96	18,35	28,32	28,66	33,24	30,53	35,24	26,75	27,44	29,88	29,83
da	43,58	24,22	18,04	17,50	16,01	21,20	21,24	19,33	20,87	23,22	19,70	22,50	20,96
fi	55,61	20,05	23,59	23,91	26,78	21,23	16,26	9,42	0,00	0,00	0,00	22,60	22,63
gn	11,21	13,11	4,65	2,01	0,00	0,00	0,00	0,00	0,00	0,00	0,00	9,04	9,04
he	91,22	47,95	47,60	46,68	42,85	35,60	29,91	31,69	0,00	0,00	0,00	49,85	49,12
hu	85,16	30,65	36,30	43,54	37,13	22,90	23,21	14,44	0,00	0,00	0,00	35,71	35,73
id	50,66	23,37	19,56	20,82	19,69	26,37	0,00	0,00	0,00	0,00	0,00	24,25	24,12
is	33,47	9,32	14,92	9,50	7,95	0,00	0,00	0,00	0,00	0,00	0,00	11,68	11,68
it	84,99	58,75	52,94	38,04	46,94	38,62	22,35	44,32	30,71	30,72	26,61	51,53	49,01
ja	100,77	53,99	50,00	40,94	45,60	42,17	27,60	47,35	59,42	50,61	41,14	51,58	50,52
ko	61,34	33,29	28,07	28,15	26,76	18,05	25,98	41,81	22,94	33,38	0,00	33,39	33,04
nl	63,13	27,85	26,25	24,06	22,84	24,26	32,87	38,19	19,16	0,00	0,00	28,98	28,83
no	53,91	21,06	18,67	19,24	12,72	23,73	22,72	19,57	43,83	31,78	0,00	21,06	20,32
ro	46,71	16,13	11,85	12,06	28,93	21,26	42,75	0,00	0,00	0,00	0,00	15,62	15,72
sv	67,82	22,39	21,88	16,38	21,28	19,04	16,91	17,51	22,60	33,20	45,86	23,26	23,00
sw	20,22	9,89	10,93	10,80	17,41	10,00	0,00	0,00	0,00	0,00	0,00	11,02	11,04
tr	38,40	23,16	25,10	21,00	23,59	22,21	24,77	27,07	22,65	21,78	28,37	24,36	24,31
zh	69,48	41,86	34,37	32,73	24,87	28,42	38,87	38,94	37,98	34,91	33,46	38,54	36,79

Table A.3: Evolution for the indicator outlinks at selection levels

A.1.3. H3.

In Table A.4 we can see how one of the indicators which obtained low results for the selected articles is this time high in the first levels. Level zero is above the average from all articles in all cases since articles containing territory names and institutions are always very prominent. This proves right the high interest in local content and explains how specialization in this knowledge loses in interest. After level two the value starts decreasing progressively until the last levels.

Lleng.	0	1	2	3	4	5	6	7	8	9	10	Ef.	Sel.
ar	23,22	16,10	10,78	9,38	9,21	6,60	7,07	8,80	8,47	8,83	0,00	13,56	11,34
ca	37,56	12,49	19,18	12,80	14,78	17,13	11,69	10,98	9,44	0,00	0,00	15,59	15,31
cs	45,26	29,28	15,60	11,06	23,95	20,64	18,56	12,37	15,33	11,76	5,44	20,60	20,62
da	36,16	14,09	15,61	18,44	20,17	47,51	27,76	17,76	12,16	17,42	5,86	15,62	21,20
fi	36,69	13,29	20,36	22,72	12,43	14,74	9,04	5,50	0,00	0,00	0,00	16,49	16,42
gn	22,44	8,33	3,02	2,46	0,00	0,00	0,00	0,00	0,00	0,00	0,00	7,17	7,17
he	65,84	31,66	29,03	29,54	24,53	20,58	9,82	5,79	0,00	0,00	0,00	32,45	31,70
hu	45,79	18,08	26,04	23,71	16,53	16,07	8,57	7,56	0,00	0,00	0,00	21,50	21,16
id	68,64	19,13	26,38	13,62	13,41	28,67	0,00	0,00	0,00	0,00	0,00	24,23	23,92
is	18,39	6,79	11,18	7,68	7,67	0,00	0,00	0,00	0,00	0,00	0,00	8,18	8,18
it	58,61	50,58	39,32	20,35	32,28	29,64	15,39	28,07	12,67	11,19	11,01	37,65	35,45
ja	89,64	43,88	39,78	22,19	28,11	23,96	19,53	36,08	125,10	62,66	32,60	40,07	38,94
ko	50,85	19,87	22,03	16,02	29,26	14,18	20,62	23,78	17,06	5,63	0,00	21,94	22,01
nl	49,03	19,75	23,20	13,73	16,65	14,10	20,26	23,15	16,89	0,00	0,00	21,79	21,44
no	25,73	11,85	14,44	14,76	8,60	13,39	7,84	6,49	46,05	32,37	0,00	13,24	12,33
ro	54,70	9,75	10,11	5,38	21,38	31,77	29,06	0,00	0,00	0,00	0,00	11,45	11,57
sv	42,13	16,57	14,91	10,19	16,64	15,01	13,18	11,78	17,51	13,27	52,24	16,52	16,43
sw	81,21	9,06	8,08	0,44	19,14	2,00	0,00	0,00	0,00	0,00	0,00	9,01	9,05
tr	37,69	13,28	24,79	16,55	14,96	13,35	17,71	18,43	13,74	16,67	83,01	18,30	19,27
zh	52,86	34,61	29,52	21,92	14,85	16,07	20,93	26,66	26,38	17,45	17,29	30,47	27,16

Table A.4: Evolution for the indicator inlinks at selection levels

In Table A.5 Inlinks from set we can see how the most inlinked levels from the set (which in this case are the effective) are the same ones from the set. This is comprehensible as they represent more important concepts and only refer to some detail concepts for aspects in the discourse of their text. Along the levels the value decreases and does not find any recovering point like in effort dimension indicators.

Lang.	0	1	2	3	4	5	6	7	8	9	10	Ef.	Sel.
ar	16,13	10,08	6,66	4,53	2,55	1,20	1,11	0,63	0,31	0,19	0,00	8,21	5,69
ca	27,69	8,08	13,23	8,56	5,98	7,04	2,97	1,35	0,22	0,00	0,00	10,55	9,54
cs	32,97	22,49	12,20	7,56	4,86	3,65	1,19	0,87	3,62	0,12	0,02	15,64	14,88
da	27,44	11,09	12,21	9,07	5,17	15,08	6,35	1,94	1,05	1,71	0,29	11,56	7,31
fi	27,94	9,65	16,63	19,08	6,16	6,25	2,51	1,25	0,00	0,00	0,00	12,62	12,51
gn	17,19	6,23	2,27	2,14	0,00	0,00	0,00	0,00	0,00	0,00	0,00	5,44	5,44
he	51,74	25,39	21,85	18,48	12,56	8,79	3,92	3,07	0,00	0,00	0,00	25,00	23,99
hu	32,53	13,07	18,52	11,73	5,63	3,15	6,55	5,78	0,00	0,00	0,00	15,05	14,41
id	27,97	12,45	12,56	9,28	2,85	5,71	0,00	0,00	0,00	0,00	0,00	13,44	13,05
is	13,24	5,14	8,40	5,69	5,86	0,00	0,00	0,00	0,00	0,00	0,00	6,14	6,14
it	34,53	31,35	25,19	13,08	9,38	5,94	2,02	3,78	0,99	1,20	0,67	23,67	20,39
ja	76,44	37,48	34,48	17,48	15,69	10,79	4,29	8,62	32,69	19,58	6,27	34,16	31,85
ko	41,13	16,45	18,32	12,19	19,66	2,48	1,68	2,61	1,77	0,88	0,00	18,06	17,81
nl	34,75	14,13	16,84	9,00	3,23	1,69	1,28	1,45	0,39	0,00	0,00	15,58	14,54
no	16,18	7,64	8,59	8,61	2,87	4,10	1,24	0,44	3,56	0,80	0,00	8,22	6,64
ro	46,65	7,97	8,31	3,68	6,36	10,68	2,56	0,00	0,00	0,00	0,00	9,42	9,40
sv	31,35	13,05	12,25	6,90	5,60	3,60	2,12	1,54	1,58	1,00	2,88	12,92	12,24
sw	73,80	7,88	7,01	0,24	0,55	0,00	0,00	0,00	0,00	0,00	0,00	7,93	7,90
tr	20,17	8,86	16,55	8,57	6,56	2,82	3,59	1,60	0,56	1,25	7,20	11,32	8,58
zh	36,42	26,34	19,94	14,11	7,65	5,33	5,22	4,01	4,17	1,39	2,08	21,47	16,27

Table A.5: Evolution for the indicator inlinks from set (effective levels) at selection levels

In Table A.6 and Table A.7 for category memberships and category memberships from set we can see how level one is the which has many more memberships. Level zero is the root for the others and act as a general recipient. The following levels decrease showing as well how having more memberships is an indicator of interest from the editors in having their content well classified. The reason categories from set obtains zero in the levels after three is because the set reference is the effective levels and thus they are distant.

Lang.	0	1	2	3	4	5	6	7	8	9	10	Ef.	Sel.
ar	1,92	3,03	2,55	2,11	2,67	2,57	2,23	2,10	1,55	3,50	0,00	2,55	2,55
ca	1,90	3,05	2,31	1,59	1,63	1,47	1,56	1,58	1,21	0,00	0,00	2,35	2,21
cs	2,32	4,20	3,49	2,38	3,95	3,46	3,98	4,28	2,43	3,37	3,26	3,46	3,48
da	2,08	3,19	2,04	1,72	1,80	1,92	1,83	1,78	1,76	1,58	1,51	2,71	2,10
fi	1,88	2,62	2,07	1,42	1,66	1,84	1,07	1,00	0,00	0,00	0,00	2,39	2,37
gn	0,91	1,09	1,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,04	1,04
he	2,60	3,60	3,26	2,40	2,68	2,46	2,01	1,72	0,00	0,00	0,00	3,33	3,27
hu	2,16	3,46	2,22	1,97	1,60	1,61	1,93	2,44	0,00	0,00	0,00	3,05	2,95
id	1,55	2,37	1,64	1,45	1,29	1,91	0,00	0,00	0,00	0,00	0,00	2,09	2,06
it	1,30	2,54	2,07	1,72	2,02	1,90	1,57	1,87	1,83	1,89	1,44	2,06	2,02
is	0,99	1,48	1,54	1,35	1,05	0,00	0,00	0,00	0,00	0,00	0,00	1,45	1,45
ja	2,61	4,48	3,38	2,75	3,01	3,46	2,73	4,72	3,69	3,17	3,55	3,81	3,75
ko	2,31	4,25	2,59	3,93	1,65	2,35	3,88	4,18	5,06	3,63	0,00	3,73	3,67
nl	2,02	2,42	2,27	1,86	2,29	2,24	2,57	1,96	1,08	0,00	0,00	2,32	2,32
no	2,05	3,97	2,98	2,73	2,44	2,44	2,54	2,51	3,83	5,80	0,00	3,50	3,24
ro	1,92	3,16	1,34	1,24	3,66	1,82	5,63	0,00	0,00	0,00	0,00	2,39	2,39
sv	2,49	4,06	2,35	2,26	2,80	3,51	2,74	2,57	3,69	1,79	2,78	3,52	3,49
sw	1,73	2,21	1,72	1,99	2,36	1,00	0,00	0,00	0,00	0,00	0,00	1,98	1,99
tr	2,21	4,47	2,35	1,69	2,42	2,44	2,51	2,86	2,63	3,00	3,61	3,29	3,11
zh	1,87	3,50	2,72	2,72	2,02	1,89	2,18	2,86	2,96	2,96	2,59	2,95	2,81

Table A.6: Evolution for the indicator category memberships at selection levels

Lang.	0	1	2	3	4	5	6	7	8	9	10	Ef.	Sel.
ar	1,20	2,00	1,67	1,41	1,50	1,56	1,15	1,16	1,19	1,00	0,00	1,68	1,60
cs	1,56	2,74	2,33	1,34	1,19	1,50	1,51	1,17	1,14	1,12	1,07	2,23	2,17
ca	1,28	2,55	1,89	1,25	1,05	0,00	0,00	0,00	0,00	0,00	0,00	1,92	1,66
da	1,74	3,02	1,98	1,66	1,73	1,84	1,71	1,62	1,65	1,53	1,16	2,57	1,98
fi	1,20	1,81	1,39	1,09	1,12	1,08	1,01	1,00	0,00	0,00	0,00	1,64	1,64
gn	0,79	1,01	1,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,99	0,99
he	2,06	3,13	2,29	1,62	1,25	1,10	1,12	1,07	0,00	0,00	0,00	2,69	2,59
hu	1,32	2,51	1,28	1,38	1,21	1,01	1,00	1,00	0,00	0,00	0,00	2,13	2,07
id	1,13	1,80	1,33	1,19	1,13	1,12	0,00	0,00	0,00	0,00	0,00	1,61	1,59
is	0,75	1,39	1,27	1,29	1,00	0,00	0,00	0,00	0,00	0,00	0,00	1,33	1,33
it	0,87	2,07	1,61	1,33	1,38	1,39	1,07	1,11	1,08	1,18	1,06	1,63	1,56
ja	1,89	3,21	2,42	1,70	1,47	1,96	1,07	1,16	1,34	1,23	1,20	2,69	2,56
ko	1,50	2,89	1,78	2,05	1,06	1,04	1,14	1,18	1,48	1,00	0,00	2,50	2,44
nl	1,19	1,93	1,76	1,26	1,11	1,19	1,24	1,08	1,02	0,00	0,00	1,80	1,75
no	0,98	2,26	1,53	1,59	1,83	1,28	1,52	1,86	1,02	1,02	0,00	1,95	1,88
ro	1,44	2,08	1,14	1,05	1,01	1,10	1,00	0,00	0,00	0,00	0,00	1,68	1,67
sv	1,58	2,09	1,66	1,23	1,44	1,52	1,41	1,07	1,10	1,15	1,13	1,91	1,88
sw	1,27	1,71	1,63	1,00	1,00	1,00	0,00	0,00	0,00	0,00	0,00	1,38	1,38
tr	1,49	3,13	1,63	1,24	1,56	1,40	1,54	2,13	2,03	1,30	1,32	2,31	2,12
zh	1,45	2,87	2,25	2,04	1,59	1,42	1,55	1,84	1,76	2,01	1,28	2,36	2,15

Table A.7: Evolution for the indicator category memberships from set at selection levels

In Table A.8 we can see how PageRank indicator shows the same oscillations from other indicators. However, PageRank value becomes much higher than the average from all articles in level zero. Since PageRank uses inlinks and the density in the graph we can understand how the indicator condense their values in the ends where accumulate more prominence and thus in the core of the selection.

Lang.	0	1	2	3	4	5
ar	1,02E-05	5,79E-06	5,36E-06	4,60E-06	4,58E-06	4,10E-06
ca	7,32E-06	2,11E-06	2,87E-06	1,92E-06	2,16E-06	2,11E-06
cs	5,38E-06	7,05E-06	3,95E-06	2,75E-06	5,57E-06	2,75E-06
da	7,80E-06	5,12E-06	4,73E-06	9,60E-06	1,19E-05	1,19E-05
fi	5,46E-06	3,79E-06	3,97E-06	5,04E-06	2,65E-06	2,92E-06
gn	0,0014547	4,61E-04	9,07E-04	4,88E-04	0	0
he	7,31E-06	7,42E-06	5,52E-06	6,56E-06	7,67E-06	7,09E-06
hu	6,52E-06	7,19E-06	7,53E-06	6,90E-06	6,51E-06	2,22E-06
id	1,20E-05	1,05E-05	8,09E-06	1,59E-05	2,13E-05	1,79E-05
is	3,64E-05	2,13E-05	1,33E-05	1,76E-05	4,73E-06	0,00E+00
it	2,25E-06	1,54E-06	1,27E-06	7,87E-07	1,43E-06	1,57E-06
ja	3,00E-06	2,63E-06	2,09E-06	1,53E-06	1,72E-06	1,26E-06
ko	7,94E-06	4,75E-06	4,92E-06	4,04E-06	9,05E-06	2,73E-06
nl	4,49E-06	2,86E-06	2,60E-06	2,17E-06	2,18E-06	2,71E-06
no	6,97E-06	3,93E-06	4,17E-06	3,81E-06	3,21E-06	4,43E-06
ro	1,93E-05	3,00E-06	3,53E-06	2,08E-06	7,72E-06	6,62E-06
sv	6,32E-06	5,96E-06	4,10E-06	3,47E-06	3,68E-06	3,07E-06
sw	1,53E-05	2,24E-05	1,41E-05	1,29E-05	7,43E-06	0,00E+00
tr	2,46E-06	3,37E-06	4,14E-06	3,54E-06	2,23E-06	2,39E-06
zh	6,93E-06	3,82E-06	3,74E-06	2,41E-06	1,96E-06	2,39E-06

ar	3,19E-06	2,31E-06	2,57E-06	1,75E-06	0	5,73E-06	5,11E-06
ca	1,40E-06	1,55E-06	1,41E-06	0	0	2,49E-06	2,39E-06
cs	3,72E-06	2,18E-06	1,38E-06	3,38E-06	2,33E-06	4,82E-06	4,79E-06
da	1,04E-05	9,10E-06	3,96E-06	5,37E-06	2,38E-06	5,68E-06	7,87E-06
fi	2,74E-06	1,01E-06	0	0	0	3,98E-06	3,96E-06
gn	0	0	0	0	0	6,75E-04	6,75E-04
he	2,36E-06	2,06E-06	0	0	0	6,84E-06	6,85E-06
hu	2,93E-06	5,56E-07	0	0	0	7,19E-06	7,12E-06
id	0	0	0	0	0	1,02E-05	1,06E-05
is	0,00E+00	0,00E+00	0,00E+00	0	0	2,11E-05	2,11E-05
it	8,01E-07	1,34E-06	6,88E-07	6,47E-07	5,34E-07	1,24E-06	1,24E-06
ja	1,32E-06	2,81E-06	6,09E-06	3,45E-06	1,85E-06	2,29E-06	2,23E-06
ko	2,18E-06	3,30E-06	2,44E-06	1,86E-06	0	4,93E-06	4,98E-06
nl	2,49E-06	1,55E-06	1,12E-06	0	0	2,83E-06	2,79E-06
no	7,50E-06	2,87E-06	3,53E-06	5,50E-06	0	4,04E-06	4,02E-06
ro	3,68E-06	0	0	0	0	3,79E-06	3,82E-06
sv	2,26E-06	1,67E-06	3,15E-06	1,23E-06	6,01E-06	5,39E-06	5,22E-06
sw	0,00E+00	0,00E+00	0,00E+00	0	0	1,63E-05	1,63E-05
tr	2,40E-06	2,19E-06	1,88E-06	3,42E-06	4,97E-06	3,49E-06	3,17E-06
zh	2,18E-06	2,25E-06	1,72E-06	1,48E-06	1,35E-06	3,57E-06	3,05E-06

Table A.8: Evolution for the indicator PageRank value at selection levels

A.1.4. H4.

In Table A.9 with edit count we can appreciate a similar oscillation having its highest value in level zero and then decreasing and increasing along the levels. However, the difference between the first levels is not as big as in effort indicators which means that edits do not equal increase in size.

Lang.	0	1	2	3	4	5	6	7	8	9	10	Ef.	Sel.
ar	19,12	17,19	18,02	16,02	15,64	14,74	15,63	12,15	15,43	9,99	0,00	17,20	16,39
ca	33,66	20,36	18,84	15,32	18,84	16,21	9,28	10,09	10,58	0	0	18,90	18,30
cs	21,48	19,45	19,23	11,58	17,62	17,45	17,65	14,51	13,81	22,42	15,59	17,85	17,82
da	30,35	21,40	19,88	25,08	29,34	29,69	30,17	24,93	21,68	24,79	18,49	21,78	25,08
fi	18,97	15,42	15,35	19,16	14,27	19,09	23,60	17,67	0,00	0,00	0,00	15,78	15,78
gn	19,86	14,47	19,39	15,48	0,00	0,00	0,00	0,00	0,00	0,00	0,00	16,47	16,47
he	34,06	29,85	27,24	27,01	26,48	21,92	19,73	24,97	0,00	0,00	0,00	29,08	28,76
hu	12,9	16,40	11,9	18,19	15,22	10,16	10,26	8,44	0,00	0,00	0,00	13,76	13,18
id	29,09	20,91	19,43	18,82	23,42	30,61	0,00	0,00	0,00	0,00	0,00	21,03	21,17
is	35,93	45,31	35,92	35,56	34,90	0,00	0,00	0,00	0,00	0,00	0,00	43,16	43,14
it	67,23	66,22	62,20	45,35	41,11	37,05	25,85	39,24	24,48	30,32	27,07	58,33	54,10
ja	19,43	16,49	15,29	14,51	19,42	14,91	11,21	20,18	22,05	17,31	12,99	15,85	15,87
ko	24,09	19,42	19,71	19,02	31,64	15,96	18,05	16,46	18,10	8,13	0,00	19,73	19,95
nl	14,54	11,34	10,47	10,03	9,52	10,48	10,72	13,89	9,16	0,00	0,00	11,19	11,14
no	19,32	15,29	14,43	16,46	14,62	12,55	11,41	10,38	16,87	15,84	0,00	15,39	14,74
ro	32,32	15,51	15,97	14,60	40,25	22,82	23,75	0,00	0,00	0,00	0,00	16,29	16,45
sv	24,79	18,48	15,02	15,19	19,01	19,09	14,13	15,93	16,73	8,99	30,13	17,75	17,75
sw	34,46	44,73	38,49	36,15	30,50	0,00	0,00	0,00	0,00	0,00	0,00	39,33	39,28
tr	61,17	46,51	61,06	47,75	40,01	28,45	40,23	37,04	30,41	34,21	45,25	50,96	45,96
zh	52,37	38,9	23,63	9,13	7,85	8,24	7,91	9,39	7,72	6,70	7,09	38,30	9,72

Table A.9: Evolution for the indicator edit count at selection levels

In Table A.10 we can see the levels evolution for the indicator of editor count. We can see how the values along levels decrease but remain very stable, which means that the number of people who contribute is almost the same but produce more (in number of bytes and outlinks) depending on how the content is semantically related to the key words. In other words, the resultant articles size rather depends on how the editors agree it should be than the number who contribute.

Lang.	0	1	2	3	4	5	6	7	8	9	10	Ef.	Sel.
ar	13,38	13,70	13,51	12,87	12,41	12,01	12,62	11,64	12,11	10,98	0,00	13,38	12,93
ca	11,84	9,24	8,72	7,31	8,54	7,25	4,86	6,07	7,31	15,29	0,00	8,60	8,39
cs	15,11	15,14	15,52	11,82	13,76	14,65	15,07	11,97	11,05	14,59	14,98	14,62	14,58
da	15,18	12,44	11,85	14,63	17,04	17,31	16,88	14,66	13,22	14,26	11,04	12,67	14,59
fi	14,85	12,91	12,77	15,94	12,04	14,69	18,47	9,58	0,00	0,00	0,00	13,15	13,14
gn	9,51	7,64	9,97	7,87	0,00	0,00	0,00	0,00	0,00	0,00	0,00	8,50	8,50
he	18,51	19,00	17,73	17,11	17,56	16,81	15,58	14,31	0,00	0,00	0,00	18,44	18,35
hu	13,23	13,64	16,64	11,67	17,54	9,45	11,57	10,33	0,00	0,00	0,00	14,01	14,20
id	18,20	15,66	14,86	14,20	18,40	16,81	0,00	0,00	0,00	0,00	0,00	15,57	15,67
is	35,65	31,48	25,11	23,21	27,86	0,00	0,00	0,00	0,00	0,00	0,00	30,01	30,01
it	26,54	28,29	26,47	21,96	19,43	19,05	15,40	19,74	13,06	16,05	16,96	25,59	24,20
ja	13,01	12,52	11,61	10,88	13,16	10,99	9,65	15,56	15,41	13,25	10,42	11,98	11,96
ko	17,19	14,83	15,52	14,26	22,07	12,48	13,57	16,38	17,16	9,25	0,00	15,09	15,22
nl	11,05	9,73	9,08	8,64	8,82	8,92	9,47	9,71	8,59	0,00	0,00	9,55	9,51
no	14,98	13,08	12,81	13,56	12,34	11,45	10,78	9,93	13,89	14,17	0,00	13,15	12,71
ro	12,27	6,90	8,53	4,65	18,16	10,97	9,56	0,00	0,00	0,00	0,00	7,61	7,68
sv	14,63	13,02	10,85	11,60	13,67	14,87	10,48	12,89	11,53	8,50	16,77	12,55	12,60
sw	30,18	35,85	34,60	31,79	19,36	0,00	0,00	0,00	0,00	0,00	0,00	33,63	33,56
tr	26,83	24,66	29,63	24,31	19,81	14,68	20,00	19,08	15,27	17,53	21,52	25,80	23,22
zh	20,50	19,15	14,52	13,68	10,71	10,61	14,85	17,46	15,81	13,08	15,44	16,27	15,45

Table A.10: Evolution for the indicator editor count at selection levels

Also, in Table A.11 we can see how the first levels have less diversity in edition. This means that even being the same or slightly more editors than other levels, there are few who are more active. And this few decide the articles will be larger, which means not only the topic of the articles (and its degree of specialization) shapes the length but also motivates a subgroup of editors to be more active.

Lang.	0	1	2	3	4	5	6	7	8	9	10	Ef.	Sel.
ar	0,72	0,72	0,73	0,71	0,74	0,73	0,73	0,70	0,73	0,74	0,00	0,72	0,72
ca	0,72	0,76	0,77	0,62	0,76	0,75	0,69	0,71	0,90	0,79	0,00	0,72	0,72
cs	0,76	0,76	0,76	0,78	0,79	0,79	0,80	0,79	0,74	0,85	0,83	0,76	0,76
da	0,79	0,80	0,82	0,80	0,81	0,81	0,80	0,82	0,82	0,82	0,85	0,80	0,81
fi	0,72	0,72	0,72	0,71	0,68	0,73	0,79	0,80	0,00	0,00	0,00	0,72	0,72
gn	0,81	0,83	0,76	0,77	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,80	0,80
he	0,71	0,74	0,74	0,75	0,76	0,77	0,75	0,66	0,00	0,00	0,00	0,74	0,74
hu	0,40	0,42	0,41	0,41	0,43	0,36	0,48	0,49	0,00	0,00	0,00	0,41	0,31
id	0,75	0,80	0,81	0,82	0,79	0,76	0,00	0,00	0,00	0,00	0,00	0,80	0,80
is	0,71	0,70	0,72	0,74	0,77	0,00	0,00	0,00	0,00	0,00	0,00	0,70	0,70
it	0,62	0,69	0,69	0,74	0,73	0,74	0,81	0,74	0,71	0,76	0,79	0,70	0,71
ja	0,67	0,68	0,68	0,67	0,67	0,68	0,67	0,70	0,68	0,69	0,69	0,68	0,68
ko	0,73	0,75	0,75	0,75	0,74	0,77	0,76	0,76	0,80	0,71	0,00	0,75	0,75
nl	0,59	0,58	0,57	0,55	0,56	0,55	0,59	0,58	0,60	0,00	0,00	0,57	0,57
no	0,67	0,68	0,68	0,70	0,70	0,65	0,65	0,59	0,69	0,69	0,00	0,68	0,68
ro	0,59	0,60	0,81	0,69	0,58	0,56	0,65	0,00	0,00	0,00	0,00	0,68	0,68
sv	0,69	0,72	0,69	0,71	0,72	0,75	0,74	0,78	0,68	0,65	0,68	0,71	0,71
sw	0,41	0,46	0,45	0,42	0,41	0,00	0,00	0,00	0,00	0,00	0,00	0,44	0,44
tr	0,72	0,76	0,76	0,76	0,77	0,79	0,76	0,77	0,78	0,77	0,74	0,76	0,76
zh	0,67	0,69	0,74	0,73	0,73	0,74	0,76	0,70	0,72	0,70	0,70	0,71	0,72

Table A.11: Evolution for the indicator diversity coefficient at selection levels

A.2. Isolation extra indicators

A.2.1. Interwiki links direction

One of the things which was not taken into account in the isolation dimension was the direction of the interwiki links. In Tables A.12 and A.13 there are two rankings for the zero level and all articles from eight language editions. We establish that the clear leader of Wikipedia language editions is English. This can be either for a will on externalization of own content by replication (local content from level zero) or also the adaptation of articles to the own language edition.

From the comparison of both tables we can observe how in level zero the difference between the first and second language is more visible than in all articles. Then we can infer how even being replicated into English edition local content does not have interest in other language editions. At the same time, languages like French or German take second places despite in some cases it is the geographic proximity language the one which has more coincident articles (Japanese in the Chinese language edition or Spanish for the Catalan).

Lang.	1a	2a	3ra	4rta	5a
ca	es (540)	en (334)	fr (195)	it (158)	de (152)
cs	en (450)	de (334)	sk (238)	nl (200)	it (189)
da	en (350)	no (244)	de (228)	sv (221)	fr (147)
it	en (1717)	fr (1207)	es (1006)	nl (946)	pt (906)
nl	en (1573)	fr (968)	de (788)	ru (491)	pt (467)
ro	en (791)	fr (367)	ru (333)	de (285)	it (240)
sv	en (1098)	de (596)	no (471)	fr (429)	fi (372)
zh	en (1654)	ja (517)	fr (438)	de (389)	es (287)

Table A.12: Ranks of interwiki links pointing to other editions in zero selection level

Lang.	1a	2a	3ra	4rta	5a
ca	en (179290)	es (150565)	fr (128029)	nl (118281)	pt (117275)
cs	en (116436)	de (96551)	fr (86886)	pl (83513)	nl (77218)
da	en (90311)	de (75317)	fr (70823)	nl (66158)	it (63257)
it	en (494619)	fr (327702)	de (284639)	nl (256151)	es (256064)
nl	en (437285)	fr (287736)	de (274218)	it (259377)	pt (236566)
ro	en (104866)	de (74407)	it (71534)	nl (70894)	ru (68715)
sv	en (272406)	de (205119)	fr (190108)	it (162972)	nl (158248)
zh	en (188263)	ja (124190)	fr (120611)	de (119139)	es (101391)

Table A.13: Ranks of interwiki links pointing to other editions in all articles from a language edition

A.3. Edition extra indicators

A.3.1. Editors/Edits by type

The edit count type broken down in Table A.14 show that which was already known: registered users edits are majority, while second place are for those made by bots and third by anonymous ips. However, comparing the composition of the three types in the selected set and in all articles from language edition we can see how in all cases edits made by users increased and generally by anonymous. Finally, the last three columns show the percentage representation of the edit count types in the set regarding all the articles.

Lang.	Perc. User Set	Perc. Anony. Set	Perc. Bot. Set	Perc. User L.Edit.	Perc. Anony. L.Edit.	Perc. Bot L.Edit.
ar	48,7	14,4	36,9	42,1	18,4	39,5
ca	64,8	13,0	22,2	46,8	6,8	46,3
cs	50,8	13,3	35,9	49,8	11,1	39,1
da	52,6	15,9	31,5	41,4	12,7	45,8
fi	51,7	13,8	34,5	53,3	18,5	28,2
gn	29,9	4,3	65,8	14,8	2,3	82,9
he	62,3	13,8	23,8	62,0	13,0	25,0
hu	45,5	15,0	39,5	58,9	7,3	33,7
id	47,8	17,1	35,1	36,7	11,4	51,9
is	39,4	12,7	47,9	38,3	8,2	53,5
it	58,4	28,0	13,6	51,7	21,6	26,7
ja	51,1	30,0	18,9	48,2	42,4	9,4
ko	48,5	14,5	36,9	54,1	17,6	28,2
nl	56,1	13,5	30,5	51,9	12,4	35,7
no	49,5	13,4	37,1	54,5	10,4	35,1
ro	60,1	13,8	26,1	42,9	9,2	47,8
sv	59,5	18,6	21,9	52,4	16,9	30,7
sw	40,3	17,7	42,0	27,5	1,8	70,7
tr	49,9	41,7	8,5	42,7	32,8	24,6
zh	71,0	23,0	6,0	63,2	16,4	20,4

Lang.	User(Set/L.Edit)	Perc. Anony. (Set/L.Edit)	Perc. Bot (Set/L.Edit)
ar	18,1	12,3	14,6
ca	24,2	33,3	8,4
cs	19,6	23,0	17,6
da	35,9	35,3	19,4
fi	11,9	9,2	15,0
gn	48,0	44,1	18,9
he	15,9	16,8	15,0
hu	7,9	20,8	12,0
id	19,4	22,4	10,1
is	66,7	10,3	58,0
it	31,2	35,8	14,0
ja	22,4	14,9	42,4
ko	17,6	16,2	25,7
nl	7,1	7,1	5,6
no	13,1	18,5	15,3
ro	34,9	37,4	13,6
sv	22,6	21,9	14,2
sw	73,7	50,31	29,9
tr	32,6	35,5	9,6
zh	36,4	45,3	9,5

Table A.14: Percentatges of edits by type

In Table A.15 we can see the editor count by type. The percenatges are very similar but in this case the editors are slightly lower which means that the users are more active. The difference between the composition of the set and all articles also enforces the users. However, the percentage they represent in the whole community in the last three columns we can see is higher. Communities like Icelandic have a 90,3% of their users contributing in local content. Also, it is important to remark that in average a 28,6 percent of anonymous editors contribute in local content.

Lang.	Perc. User Set	Perc. Anony. Set	Perc. Bot. Set	Perc. User L.Edit.	Perc. Anony. L.Edit.	Perc. Bot L.Edit.
ar	39,3	18,8	41,9	32,1	22,9	45,0
ca	52,0	16,2	31,8	36,6	8,1	55,3
cs	42,4	17,5	40,0	45,7	11,7	42,6
da	45,5	16,7	37,8	38,0	13,5	48,6
fi	41,5	18,0	40,5	46,9	19,2	33,9
gn	28,7	5,1	66,1	18,7	3,6	77,7
he	47,9	19,7	32,5	54,2	18,3	27,5
hu	44,6	25,7	28,6	130,5	9,8	46,9
id	39,6	20,1	40,3	32,7	12,9	54,4
is	37,9	20,1	42,0	39,1	10,6	50,3
it	44,4	33,5	22,1	38,6	23,8	37,7
ja	42,5	29,7	27,7	42,6	45,1	12,3
ko	39,6	17,9	42,6	45,0	20,2	34,8
nl	43,6	18,1	38,2	42,7	13,3	44,0
no	40,2	17,9	41,9	45,1	11,6	43,3
ro	45,9	16,8	37,2	31,5	10,5	58,0
sv	49,5	21,8	28,7	44,7	18,3	37,0
sw	42,9	28,8	28,3	25,3	2,8	71,9
tr	37,4	51,2	11,4	34,2	38,0	27,8
zh	57,4	31,8	10,8	48,7	21,9	29,4

Lang.	User(Set/L.Edit)	Perc. Anony. (Set/L.Edit)	Perc. Bot (Set/L.Edit)
ar	30,3	20,2	23,0
ca	21,6	30,3	8,7
cs	25,5	41,0	25,8
da	33,6	34,9	21,9
fi	16,0	17,0	21,6
gn	42,2	39,1	23,3
he	18,3	22,3	24,6
hu	7,9	60,6	14,0
id	24,5	31,6	15,0
is	90,3	17,7	77,9
it	27,3	33,5	13,9
ja	28,1	18,6	63,6
ko	28,5	28,6	39,7
nl	9,8	13,0	8,3
no	20,2	35,0	21,9
ro	31,4	34,3	13,8
sv	26,4	28,3	18,5
sw	157,8	96,94	36,7
tr	29,3	36,2	11,0
zh	35,7	44,0	11,2

Table A.15: Percentatges of editors by type

A.3.2. Group diversity coefficient

The use of the same concept of diversity coefficient can be applied to the selected set and all articles. Instead of ranking the editors from one article by number of edits in the same, we can rank the whole set and all the editors which edit in each. This allow us calculating the diversity coefficient by different percentages of edits. In this case we choose [0,100] discretized by 2,5. From this we obtain a curve which is the power law already known.

However, our estimation was that the set of articles would obtain lower diversity coefficients than all articles. But it has not been confirmed. We checked the difference between the curves in each languages and in only one case there has been more diversity in all articles. For this we understand that there are many bot edits in the top of all articles ranking than in set ranking.

The obtained values in Table A.16 are negative in all eight cases but the Catalan language. The final indicator value is the subtraction of both curves (which leaves an area in between) in relation to the set curve. As a future test it would be interesting to try the same coefficient limited to edits and editors which are registred users. In Figure A.1 we can see the two curves in Catalan language edition.

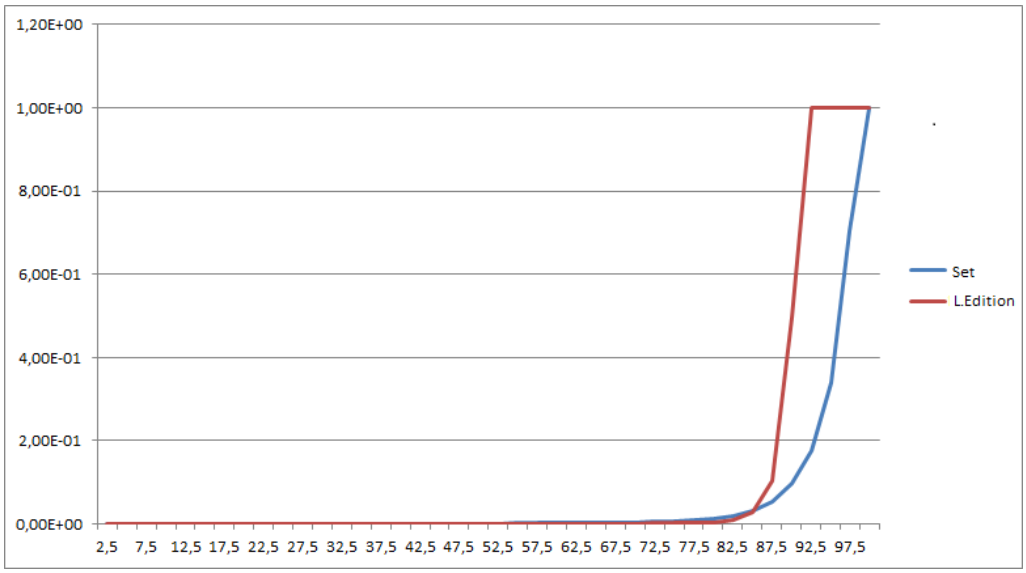

Figure A.1: Evolution of diversity coefficient on the set and language edition

Language	Sum	Area	Indicator Value	Pos.
ca	2.486	1.2	48.731	1
cs	2.770	-1.0	-36.105	7
da	2.893	-1.0	-34.586	5
it	3.379	-1.9	-55.565	8
nl	2.794	-1.0	-35.215	6
ro	2.440	-0.7	-30.021	3
sv	3.172	-1.1	-34.337	4
zh	3.087	-0.9	-28.554	2

Table A.16: Result for group diversity coefficient

A.4. Temporal extra indicators

A.4.1. Interest fluctuation

Another way of understanding the growth in sets of articles is evaluating its stability. If the interest is at the same level as the language edition, the percentage of creation in the set of articles will be similar to the one in all articles. In Table A.17 we calculated the percentages for each semester since 2002 until last month and generated two calculations with the sum of differences between periods and the standard deviation.

Thus the differences enable us to see if there are abrupt changes in the creation rate, while the standard deviation shows the stability. We can see how, out of the eight language editions in which we tried both calculations, the Catalan language edition is not as stable as other language editions but its changes are not abrupt. However, Czech or Danish are much more abrupt but more stables. The most abrupt and least stable is the Romanian.

Llengües	200201	200202	200301	200302	200401	200402	200501	200502	200601
ca	1,8	0,0	7,9	14,7	56,3	23,0	19,8	12,3	10,5
cs	0,0	0,0	11,6	40,9	11,6	17,6	14,2	25,7	21,5
da	54,1	18,8	18,3	19,5	20,1	24,0	28,9	47,0	35,1
it	11,6	8,8	18,1	19,6	23	51,3	20	9,8	17,7
nl	11,2	10,4	14,0	23,2	16,9	17,5	13,6	19,3	9,1
ro	0,0	0,0	0,0	4,7	17,0	7,6	18,8	30,2	77,9
sv	23,5	42,8	29,9	15,8	21,6	31,0	30,6	25,0	27,9
zh	0,0	32,5	29,3	27,8	23,1	28,4	33,4	33,2	34,2

200602	200701	200702	200801	200802	200901	200902	201001	201002	**Diff.**	**Standard Dev.**
12,5	14,5	18,9	13,9	16,7	13,4	16,7	17,5	9,3	138,8	11,4
37,9	19,8	19,7	17,9	20,3	22,7	20,8	43,3	22,7	172,5	9,3
33,7	33,0	25,8	29,4	24,6	24,5	35,1	34,4	34,3	105,2	9,3
18,2	13,8	15,8	15,7	14,9	12,7	9,5	9,8	11	111,6	9,3
17,1	17,7	12,1	12,6	13,8	20,6	18,2	23,7	9,6	91,92	4,3
30,7	42,1	11,6	23,4	31,0	25,8	21,7	22,7	34,4	227,68	16,8
27,9	25,5	30,0	30,1	26,5	29,0	28,8	25,9	29,1	92,33	5,2
29,3	27,0	21,6	22,2	24,7	14,5	20,0	31,6	26,6	79,3	5,2

Table A.17: Relative value for growth of Set regarding the language edition growth

A.4.2. Antiquity based predictor

Last, we want understand when and in which percentage an important growth is produced in each level of selection. We understand that the antiquity may understand us the current state of the local content and therefore predict its possible growth. Any encyclopedia can increase its number of articles indefinitely but the selected set will be limited to the required notability of topics which legitimates an article.

In Table A.18 we can see eight language editions which we selected out of the twenty. Most of the levels zero have their highest period of articles creation in 2006-2007 but Chinese and Italian. The following levels are every one newer, since the topical specialization which they provide may come usually after the more general articles. This is confirmed by at least the fourth first levels in almost all languages as a trend.

Llengües	0	1	2	3	4
ca	200702 (15.53)	200702 (16.61)	200902 (13.46)	201001 (19.37)	201001 (19.3)
cs	200602 (11.72)	200602 (13.71)	200602 (26.39)	201001 (48.33)	200701 (12.57)
da	200701 (10.57)	200902 (9.38)	200502 (29.62)	201001 (12.56)	200702 (8.79)
it	201001 (12.12)	200602 (12.04)	200801 (11.86)	200402 (15.53)	200602 (17.52)
nl	200702 (9.06)	200701 (10.97)	200702 (10.02)	201001 (14.38)	200701 (15.66)
ro	200602 (17.47)	200701 (32.29)	200601 (75.57)	200702 (37.3)	200702 (16.99)
sv	200701 (12.21)	200601 (11.15)	200601 (10.98)	200402 (12.29)	200601 (14.21)
zh	200901 (10.03)	201001 (11.69)	200602 (12.27)	200602 (11.73)	200902 (14.7)

5	6	7	8	9	10
201001 (37.31)	201001 (48.89)	201001 (21.19)	200902 (45.93)	200702 (57.14)	0
200801 (24.39)	200801 (23.94)	200701 (18.98)	200701 (26.66)	200801 (23.85)	200602 (25.92)
200602 (8.88)	200502 (11.15)	200802 (10.53)	200802 (14.31)	200901 (12.57)	200801 (18.23)
200502 (18.5)	200702 (36.08)	200602 (14.26)	201002 (21.31)	200801 (18.79)	200601 (43.62)
201001 (11.29)	200602 (12.75)	200502 (13.91)	200502 (27.84)	0	0
200602 (45.25)	200401 (33.33)	0	0	0	0
200602 (12.39)	200601 (25.41)	200601 (31.26)	200901 (40.9)	200901 (87.8)	0
200902 (18.6)	200902 (15.75)	200702 (11.11)	200702 (12.58)	200702 (13.79)	200801 (13.04)

Table A.18: Most growth period by level

APPENDIX B. WEB REPRESENTATION

In this appendix we can find two screen captures of the webpage which reproduces this same document. This was a petition of different Wikipedia language edition comunities, which were interested in having the results of this project. In Figure B.1 and B.2 there is the abstract and the autoreferentiality dimensions sections. In Figure B.3 and B.4 there is the wikAPIdia and Index sections.

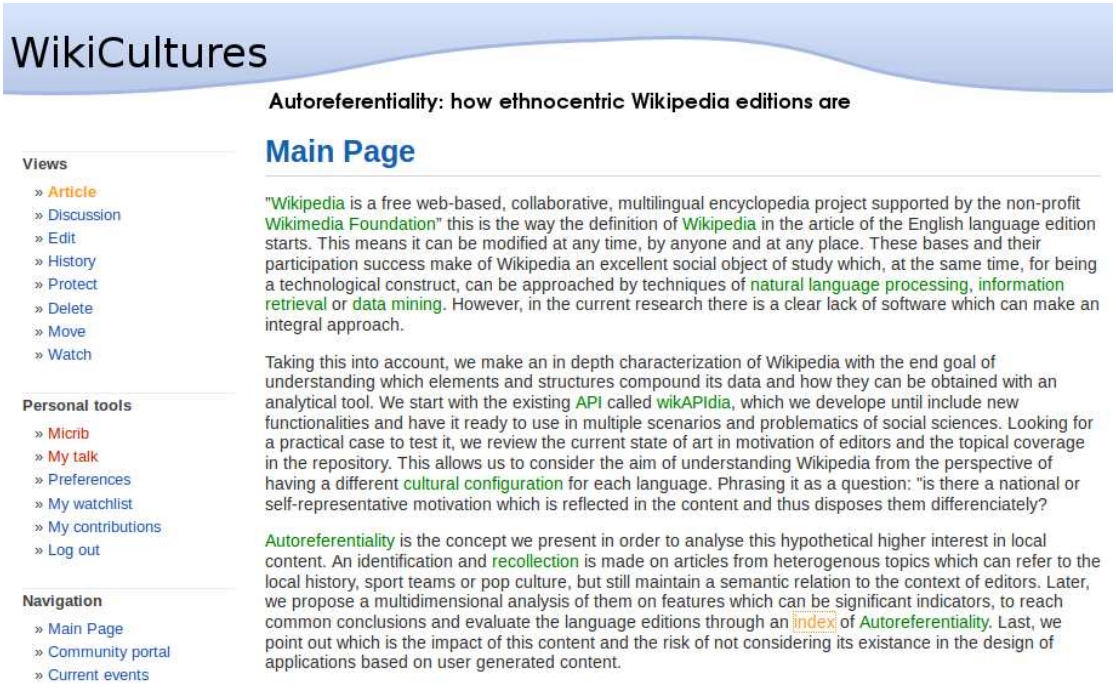

Figure B.1: Capture screen from 'Abstract' and Main page of 'WikiCultures'

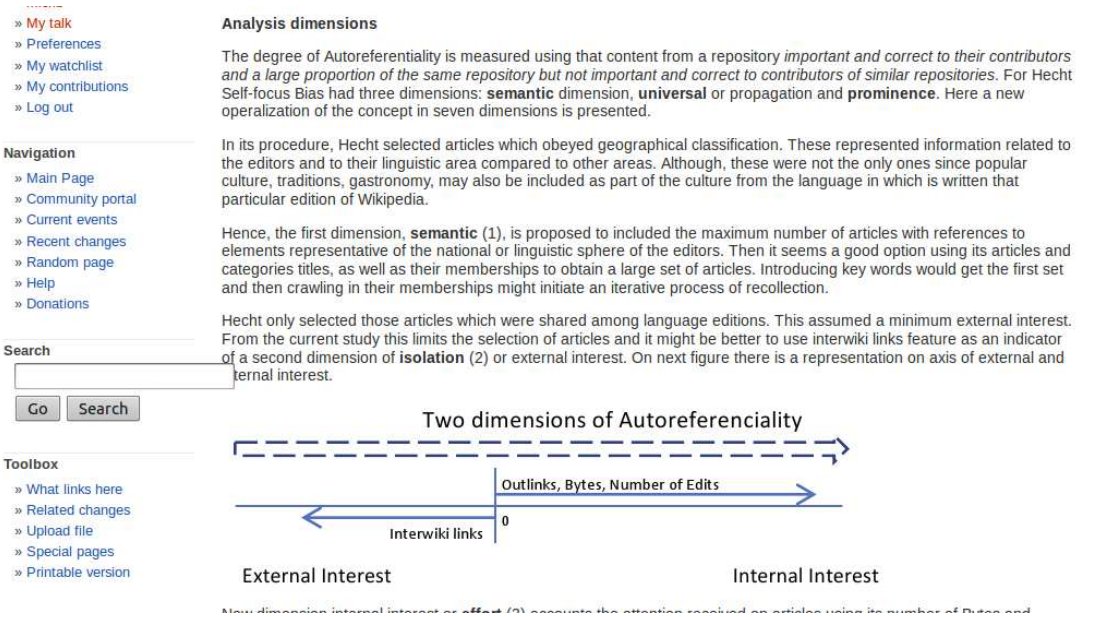

Figure B.2: Capture screen from section 'Autoreferentiality' of 'WikiCultures'

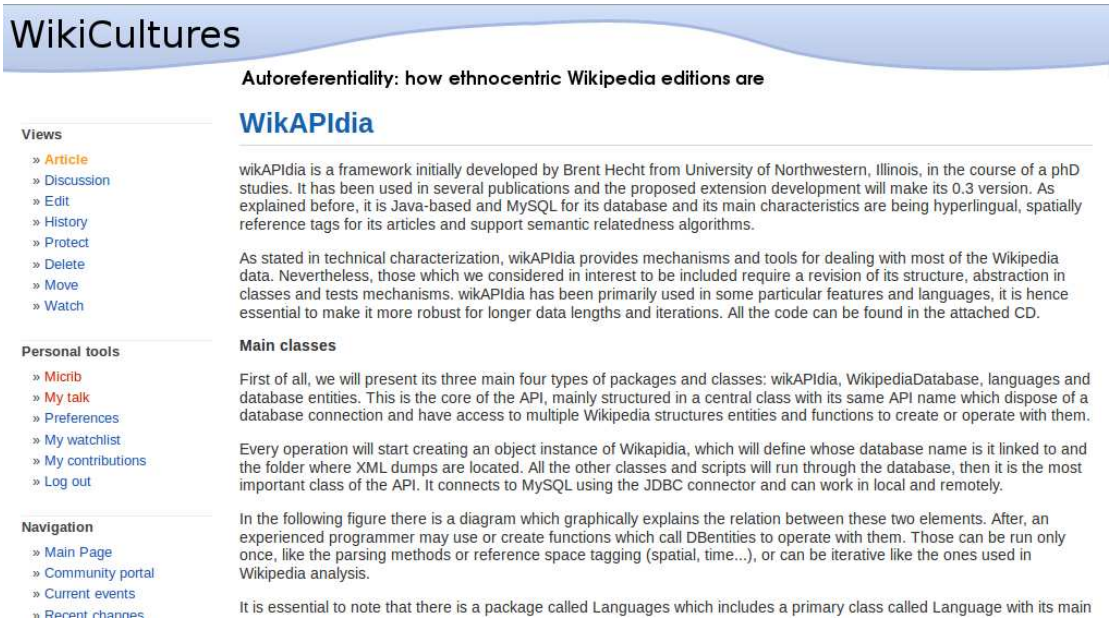

Figure B.3: Capture screen from section 'WikAPIdia' of 'WikiCultures'

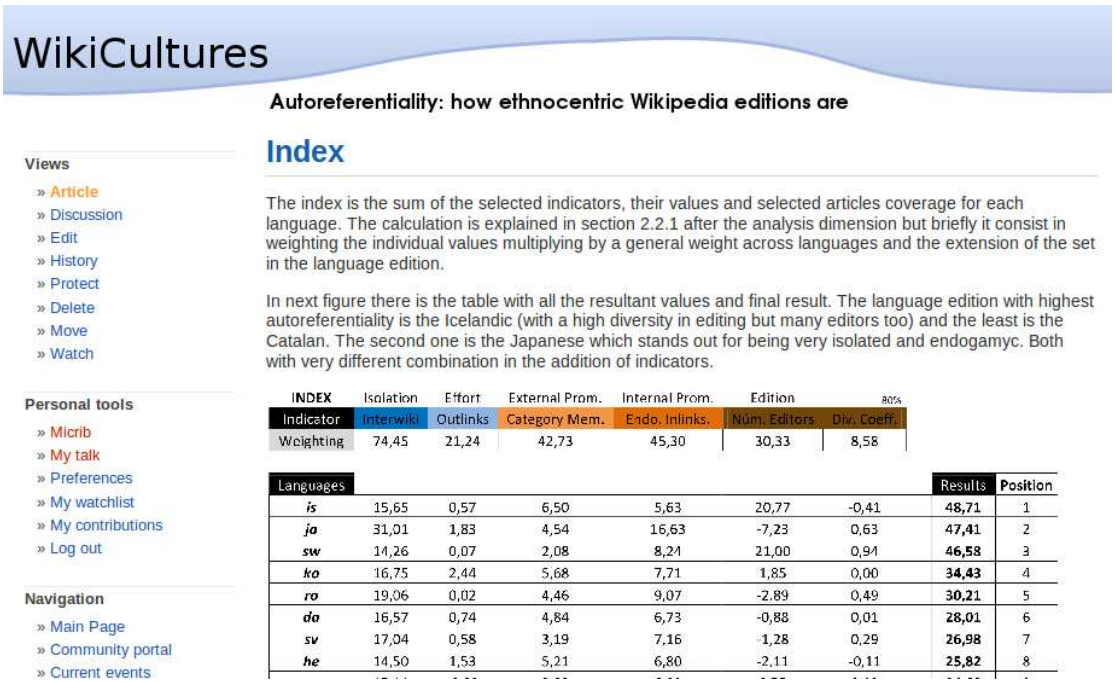

Figure B.4: Capture screen from section 'Index' results of 'WikiCultures'

The wiki format is also a way of encoraging the community to keep working on this direction and close the circle. What is studied in a wiki is then explained in a wiki.