BORDERS. Visual analysis of Cinema's inner dynamics and evolutions. A case study based on the Internet Movie Database

Giovanni MAGNI[*a], Paolo CIUCCARELLI [a], Giorgio UBOLDI [a] and Giorgio CAVIGLIA [b]

[a] Politecnico di Milano; [b] Stanford University

Since the launch of the Internet Movie Database (abbreviated IMDb), in 1990, a large amount of data about movies has been collected online, giving us the possibility to analyse the film industry on a large scale. This paper describes the design and some results of 'Borders. A geo-political atlas of film's production', a project that aims at visually represent and exploring the multi-dimensional and heterogeneous data coming from this online archive. The main goal is to visually analyse the relationships between geo-political regions and countries, using the data related to the film industry as a privileged point of view.

***Keywords**: Cinema; data visualization; online archives; networks; internet movie database*

Introduction

The opening of enormous online databases and the increasing availability of tools to extract and analyse huge amount of data are providing a unique opportunity for researchers and scholars to represent and study social and cultural complex systems.

The work presented here aims at making a step in the direction of deepening the possibilities offered by the abundance of online data to define new modes of observation, exploration, and interpretation to analyse the evolution of a large scale social and cultural phenomenon such as the film industry.

The film industry is a cultural and economical system that has grown enough to develop proper dynamics thanks to phenomena such as co-

[*] Corresponding author: **Giovanni Magni** | e-mail: **giovanni1.magni@mail.polimi.it**

production and tax-incentives (Morawetz and Norbert, 2007), therefore it appears as a world extremely connected to reality and history but with its own inner geographies, built and shaped during time, film by film.

The birth of digital archives and online platforms to collect, store and share information about the world of cinema (e.g.:IMDb, Rotten Tomatoes, Wikipedia) opened new possibilities to study and analyse the film production as a collective system rather than a sum of individual cases.

In our work we focused on the quantitative and visual analysis of this data as a key to understand the evolution of large scale phenomena such as the relationship between countries and different geo-political and cultural areas within the global film production.

The goal is to offer a macroscopic perspective both on the studied subject (whole national identities rather than small social groups) and for what concern typology of used data (analyzing the whole dataset available rather than rely on small subsets).

In this perspective, information and data visualizations have increasingly emerged as essential tools to explore and make sense out of the growing quantity and variety of available data, playing a key role in research activity and not just the final output (Jessop, 2008).

The paper collects and discusses the whole process that has lead to the definition and the development of a thematic atlas, namely 'Borders. A geo-political atlas of film production', starting from the theoretical and the methodological assumptions behind the collection and transformation of the data to the use of network visualizations to produce an image of the phenomenon.

The project is part of a Master of Science thesis in Communication Design and it has been developed in collaboration with the DensityDesign Research Lab. (densitydesign.org) at Politecnico di Milano.

As done in thematic atlases (e.g. Historical) where the world's geography is modified in order to depict the evolving political landscape or a specific phenomenon, 'Borders' aims to analyse quantitatively the evolution of global film production during time showing how data collected online could be an interesting point of view on more complex phenomena. The final output of the thesis is a visual artefact in which we represent connections between countries and cultures involved in the production of movies, using metadata coming from online database, in particular IMDb and Wikipedia.

Background

A preliminary research showed us how even if there are many projects on this theme (especially outside academic environment) there was a lack of projects and analysis of cinema able to consider large amount of data.

Except for some interesting projects made on similar datasets (Herr et al., 2007; Ahmed et al., 2007) usually this kind of works focuses on sample surveys, on the production of a single director or on an artistic movement.

At the same time most of the analysis considered are related to technical features of films, many projects try to see movies from a different point of view. They focus on various aspects of the productions trying to generate for example chromatic overviews ('CinemaRedux', Dawes, 2004), to understand features of editing ('Visualizing Vertov', Manovich, 2013), network of characters ('Lostalgic', Ortiz) or trying to correlate different movies to each other using properties such as genre, director or actors involved, rarely considering the organization's system that exists behind every single production.

Therefore, the aim is not to study cinema from a 'technical' point of view but rather to give it a social role, to use it as a reading key to study and highlight historical and social dynamics developed during time, in this paper we would like to show potentialities of this project showing results coming directly from the Master of Science thesis 'Borders'.

Methodology

The starting point of our analysis is the single production, whether it is a movie, a short movie, a TV series or an episode.

We can retrieve a large amount of information about various aspects for each production, for example we can easily know the people who worked in or which companies have been involved. Most of these data can be related to a geographic position, a spatiality or a national belonging and this, from our point of view, creates connections between the nationality of the production itself and the countries that appear in the data.

The idea was to analyse these connections counting how many times they occurred during years, giving them a proper dimension and using them to graphically represent a social and economic evolution of collaborations during time.

The process we put in place is basically simple, we went through the archives we had and movie by movie we kept trace of every link between

the nationality of the production and any other country that, for various reasons, is releated to the production and consequently appears in the data.

This way we have been able to give to these connections a size, a dimension, which has been use lately to create visual representations of the various phenomena treated, we were not interested in why these correlations happen, we wanted to know how many times they happen.

Data Collection and Process

Since the target of the project was to create a quantitatively significant analysis, we could not base the whole project on a small amount of data, so we would have to rely on the biggest archives available online.

The data for this project comes mainly from two sources, the Internet Movie Database (imdb.com) and Wikipedia (wikipedia.org).

IMDb is a free online database of information related to movies, Tv series, TV programs and video games that takes in account many metadata related to the productions and the contents.

Since we wanted to analyse the geo-political and cultural dimension of the system, the first step to obtain a consistent database to work on, was to extract all the movies' features that, as we said before, can give us connections between countries.

For what concern IMDb, we started from a subset of the whole archive (imdb.com/interfaces) focusing on the most complete and well recorded metadata, specifically:

- Locations of shooting (774.687 locations known)
- Companies involved into production (1.632.046 collaborations known)
- Languages spoken inside of movies (932.943 language appearences known)
- Release dates (1.008.384 dates known)

To have a more complete database to work on we also decided to integrate the information coming from IMDb with other data taken from different linguistic versions of Wikipedia (wikipedia.org), the most famous and collaboratively edited, multilingual, free Internet encyclopedia.

Regarding these data, we have analysed 160.031 different pages related to films on the various language versions of the website (285 at the last official update) and their own detail level (page weight).

For all the data cleaning and manipulating we personally created scripts in the Python programming language which revealed itself as very powerful handling huge text files.

The datasets we obtained from the websites were very raw, therefore the first important step has been a concrete cleaning to remove all the useless information that were incomplete, inconsistent or secondary.

After that we proceeded to count, as we said previously, the amount of connections inside of the archives obtaining a list for each aspect considered, structured as the following extract:

Table 1 Example of list of connections between countries.

Country of production	Country involved	Amount of occurrences
…	…	…
USA	Spain	1222
USA	State of Palestine	8
USA	Sri Lanka	21
…	…	…

These values have been lately used to quantify the 'attraction' between different countries and to develop visual representations of the subject.

Many times happened that a single movie has two or more nationalities since the production companies could come from different nations, in these cases we considered all of them and we repeated the process for each nationality.

Visualizations

Generally, one of the best ways to represent large scale social dynamics and relationships is the network visualization. Networks allow us to shape and to explore a phenomenon, to focus on clusters and, using quantitative methods for measuring social relationships (e.g. degree distribution, clustering, closeness, centrality, distance), to understand how they interact with each other inside of the graph.

Networks are, such as Information Visualization in general, rather than the last step of a process of visual representation, a tool used to understand and to explore a subject and to see the structure of a large archive (Van Ham et al., 2009).

Using a software package such as Gephi (gephi.org), an open source network analysis application which allows us to visualize and to explore large graphs coming from complex datasets (Bastian et al., 2009), the values

obtained from metadata of individual films (the list we saw before) have been used as a base on which build networks of countries considering them as forces of attraction between nodes of the networks.

Using different positioning layouts such as the force directed Force Atlas 2 (Bastian et al., 2011) we were able to produce different kind of visualizations. The main idea was to explore the potentiality of the network visualization in order to analyse the relationships between countries based on the data coming from our dataset, ignoring their actual geographical position.

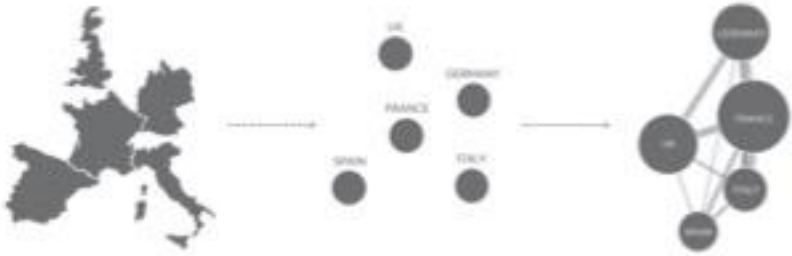

Figure 1 How it works; comparison between real geography and positioning of same countries in a world created by film productions.

Results

'Borders' is the result of a process of harvesting, analysis and visual processing of data, it consists of five chapters, each one of them refers to a different dimension of the information we analysed.

Explaining every single visualization obtained would take too long, therefore we are showing just some of the most interesting examples of the results.

1. Location's Analysis

'Where' a shot is taken is a choice that depends on various causes, two of them are costs of production and the requirement to move to a specific place according to the film's plot. An entire cast moves to a different location to follow the film's theme which can require specific place and sets or to save on production's costs moving to places where, for multiple reasons, it results cheaper.

Analysing the whole list of locations recorded on IMDB, the aim is to visualize which are the countries that take advantage from these dynamics

and how nations behave differently in this process of import/export of shooting.

In figure 02 we can see how nodes inside of the general network acquire more or less authority bases on which parameter we choose to define their size. Difference is between this two parameters:

- inDegree: the bigger is the amount of countries which made locations into the referred node, the bigger is the node.
- outDegree: the bigger is the amount of countries in which movies of the referred node made locations, the bigger is the node.

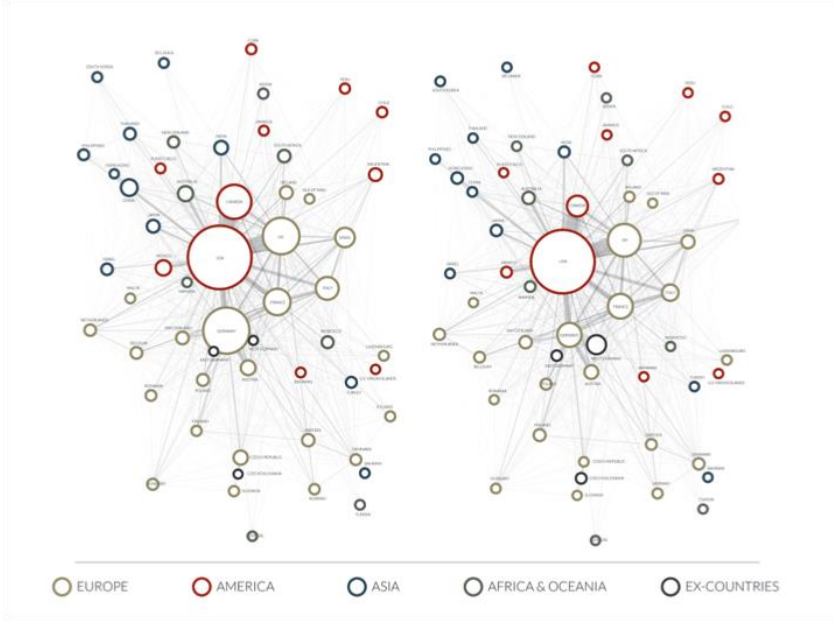

Figure 2 Network of countries linked by amount of location. Two nodes are close if the movies made in one of them have done a lot of shooting in the other one. Comparison between different node's size parameters, inDegree (left) and outDegree (right).

Using the same node's positioning, we can modify the visual substance of each country inside of the representation and we can see how there are big dissimilarity between the two variations.

At the same time, using the same information, an additional analysis on individual countries can be done, we can visualize the percentage of

locations made in a foreign country related to the total amount of locations recorded in the archive and see how different nations behave differently.

We can see in figure 03 some examples:

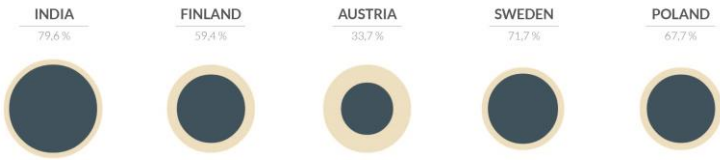

Figure 3 Example of individual country profiling. The light big circle is proportioned to the amount of locations known, the blue one represents the amount of location within the country.

2. Companies Analysis

A study on collaborations between national productions and different companies shows again a sort of economical side of this world. The most interesting part of this analysis is made by a network of countries more or less attracted to each other according to a value which is a count of times that a particular connection occurred (for example amount of times that Italian movies involved Spanish companies). As we see in figure 04 this network is dominated by western and economically better developed countries, it basically shows importance of a national film's industry within the global production.

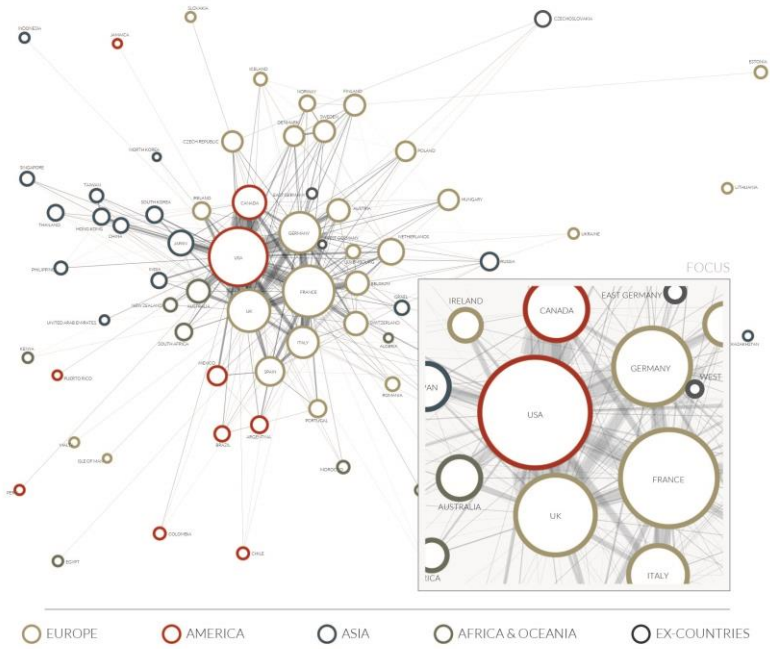

Figure 4 Network of countries linked by amount of companies involved in respective productions. The closer two nodes are, the bigger is the amount of times that the collaboration between elements of the two nodes appears in the archives.

At the same time it's interesting to focus on smaller economic systems and geographic areas, showing the historical evolution of inner dynamics.

In figure 05 we can see how the situation in the European continent has evolved and strongly changed during time, the amount of connections has clearly increased according to the dense network of connections between nodes and the amount of countries of which we have available data in the archive.

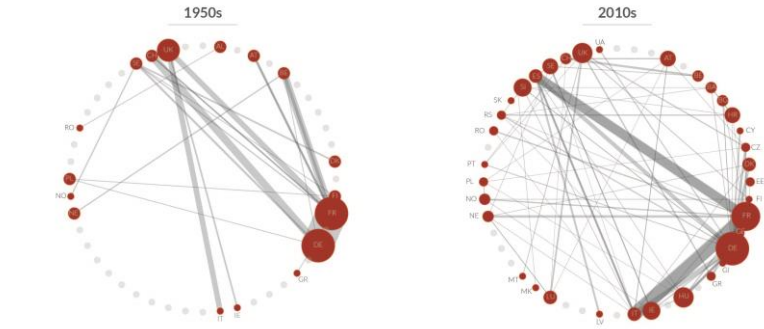

Figure 5 Circular network of european countries linked by amount of companies involved in respective productions. Comparison between 1950s and 2010s. Values are normalized to 2010s connections.

As we did before, we can visualize (figure 06) how an individual country evolved its own production involving foreign companies, showing interesting trends and a much more detailed point of view.

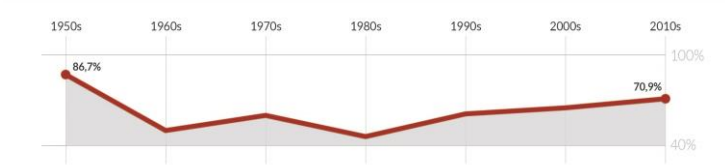

Figure 6 Percentage of Canadian companies involved in the Canadian movie's production during time on the total amount of companies recorded.

3. Languages Analysis

As we said previously, a kind of information available on IMDB concerns languages of dialogs within the movies. Our opinion was that themes debated within a national film's production are strongly connected to the history of the country and to events in which the nation itself has been involved in. Therefore a strong appearance of a foreign language in the

movies' dialogs of a specific country could represent a sort of link, a connection between different cultures and nations considered.

A bipartite network show us in figure 07 how countries and languages arrange themselves mutually, according to connections between them, generating new clusters and showing relationship developed during time. It's important to point out that, to highlight this feature, within the network has not been considered the link between a nation and its own mother language, obviously this value is numerically much bigger than any other connection and should force the network into a not interesting shape.

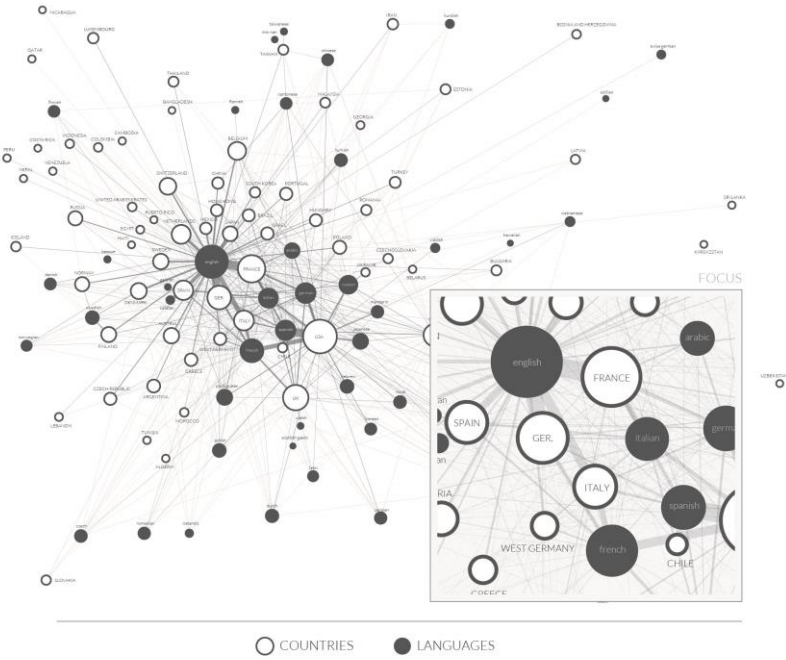

Figure 7 Bipartite network of countries and languages, nodes are linked by amount of times that a language appears in the movies of a country therefore a country and a language are close if the idiom appears a lot of times in the productions of that nation.

4. Release Dates Analysis

In this case, available data revealed itself as messy and confusing compared to the previous ones, tracking release dates of movies in different countries is not easy and it shows another peculiarity, in the IMDB archive we can find complete data regarding most famous and biggest productions but at the same time, data regarding small national systems and less important movies are incomplete or not significant.

To develop a correct analysis of the global movies' distribution phenomenon it was necessary to take a step back and base it on a reliable set of data. Specifically we decided to focus and analyse distribution of US movies around the world, indeed into the database they are quantitatively much more represented than the other countries and related release dates are better recorded.

Furthermore we decided not to evaluate data related to TV programs and TV series, which follows different and specific ways of distribution.

We thought that the better way to verify potential trends during time of this particular aspect was to visualize in each decade how many American movies were released in any other nation and how far (days of delay) from the American release date, generating a sort of economic and cultural detachment between United States (which can be considerate as leading country) and any other country.

Supposition is that a movie is released earlier where there is more interest and therefore more chance to get a gain from it.

In figure 08 we have a clear representation of the phenomenon and how it evolved during time, speeding up. For example, if we look at two different decades, the nineties and the last one from 2010 and beyond, we can see how there are significant differences between them. In the first case, American movies were not released before a six months delay from the American date, in the last years they need no more than ninety days to get in almost every country, witnessing the process of globalization, expansion of the Hollywood industry and technology evolution of the last century.

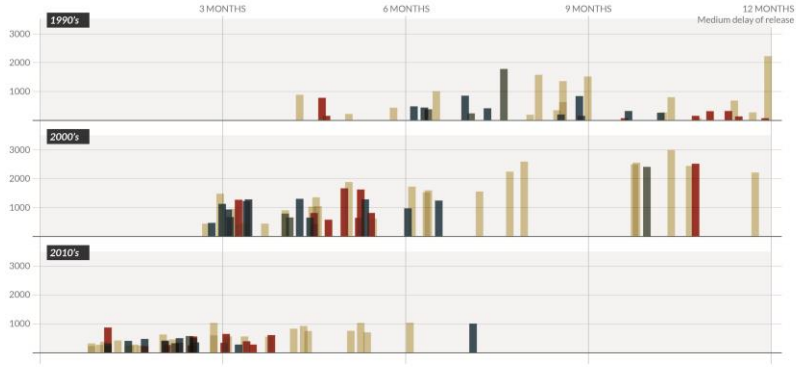

Figure 8 Evolution of American movies distribution during time. Each bar represent a country colored by continent, proportioned to amount of movies distributed and moved to the right depending on medium delay from the American release date.

5. Wikipedia's analysis

What we did in this last paragraph was to verify how films of each country are represented on the different Wikipedian linguistic versions through related pages.

To collect necessary data we used both DbPedia (dbpedia.org) and the encyclopedia's APIs.

DbPedia is a crowd-sourced community which extract structured information from Wikipedia, we relied on it to obtain a list movies with all the respective pages on the different linguistic versions of the online encyclopedia. What we wanted to do was to verify the overall interest on national productions evaluating their amount of pages on each Wikipedia.

Since pages are different between each other, we had to find a way to evaluate the importance of every single one, the idea was that a page full of information could not be considered likewise a page which is just sketched or incomplete.

To overcome this problem we decided to use the 'Page size' data which is available for each one.

We were not interested in the accuracy of information within the pages which can be problematic and untrustworthy (Kittur *et al.*, 2008), otherwise since we just wanted to consider differences between pages, the weight (expressed in bytes) was enough since we did not care about kind of

informations, indeed in any case lot of data means lot of interest which was what we were searching for.

So, the value which connects countries and Wikipedias in figure 09 has been calculated not just counting the amount of pages but summing their weight.

Our supposition was that people are more inclined to upload data on their own language version of the encyclopedia and so what we got is a bipartite network in which we can see how national movie's production is available on different Wikipedias.

A problem here was how to consider the English version, since it is a kind of global web landmark which we could not relate to a language culture, moreover since we use it as a starting point where to find list of movies categorized by country of origin (dpPedia.org refers to the english version of the encyclopedia), it would result as a version where all the pages related to movies are available, so we removed it and we considered it as we did with mother languages in the paragraphs before.

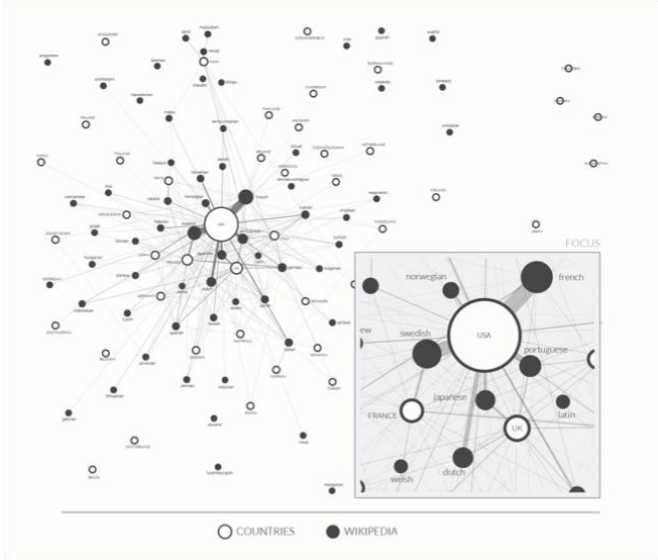

Figure 9 Bipartite network of countries and Wikipedias. Nodes are linked by total amount of page sizes related to movies of each country available on the encyclopedias, therefore a proximity between a nation and a website means that the national production is well recorded on that particular Wikipedia.

Discussion and future work

The enormous starting database gave us big hints but at the same time big doubts on the actual feasibility of this analysis, while human eye is able to evaluate in advance and wonders a certain kind of results on small datasets, when we deal with archives as big as what we had we can only try some sample surveys to explore and verify the actual consistency of the used information.

Risk to obtain non significant or non consistent visualizations was big but, after all, we saw how data have revealed themselves as graphically analyzable, showing trends and evolutions during time. If we consider networks, one of the possibilities was to obtain unclear or scattered results which could be very hard to understand, on the contrary countries grouped into clusters easing the comprehension of visualizations.

Visual representations return clear images of a non tangible world with new geographies built up film by film and year by year. The world of cinema is a complex system characterised by no accidental dynamics which have been developed during time.

The visualizations we obtained do not solve this complexity, on the opposite they try to preserve it creating a way to show inner relations and roles of individual national productions, revealing that connections that would otherwise remain hidden inside the archives without any chance to get explored and understood.

Getting access to this huge amount of data, combined with the simplification of processes of visual exploration given us by new software and by recent general interest in data visualization, gives us the chance to verify different features of different phenomena such as Cinema or Literature, without relying on sample surveys.

Results show, especially where is possible and reasonable to verify an historical evolution, various aspects related to different sides of society.

They concern economical features when we analyse locations and companies involved into productions, otherwise they concern some more social-cultural features when we focus on release dates, languages or the Wikipedian part of the subject. What we mean is that with these archives of information we can obtain a complete and variegated picture of global situation and dynamics evolved during the last century.

We want to highlight that in this project has been used just a little part of data available on IMDB. In future could be very interesting to aggregate new datasets extending the research and showing new aspects of the film

industry. A study on the social networks between actors, directors and other people involved in the production or on the budgets and incomes in different geographic areas are just two of the various chances offered by IMDb which can lead to a future bigger research and projects.

At the same time one of the weaknesses of this project is its strong connection with the user generated nature of the data. It is important to remind that the whole analysis does not come from an official archive but on the user generated content available on these platforms.

These websites have both positive and negative features. First of all during time they became a web landmark, attracting to themselves more and more users and data, generating an enormous archive on which we can develop analysis. Secondly this kind of information represent an image of the global interest of people in cinema rather than an 'official record' of the whole movie's production, since the upload of data is assigned to the users the archive reflects the matter of movies into the collective social imaginary (bigger the fame of a film is, bigger is the chance of having information into the database).

These websites have some lack for what concern the kind of users involved, national production such as Indian and Chinese (the first one much more bigger than the second one) are not well reported, both for a possible lack of interest by indian/chinese users (that maybe use to refer to other websites) and for a shortage of interest by western users to that kind of national productions that are numerically considerable, especially for the Indian one. To avoid incorrect or incomplete observations it's therefore necessary to keep in mind this peculiarities of the global society that affect participation and completeness of these archives.

However it could be very interesting to do the same analysis we did on IMDb starting from official databases and to verify differences and equivalences in these two images of the same world coming from various sources, to see disparities between the 'official' or verified situation and the collective interest in cinema.

'Borders' is the results of a Master of Science degree in Communication Design, therefore it has been developed since the beginning with a Design point of view.

We think that a collaboration with domain experts such as cinema's critics, data journalists, analysts or sociologists could provide more interesting insights and a deeper analysis about the subject.

This debate, which lacked during the project, could give us that interpretation of results that, being designers, we can't have the experience and skills to do.

What we create is an interpretation of the data, explored and graphically elaborated various times in that design process which lead us back and forth, from information to visualization repeatedly, in a continuous exchange of hints and elaborations.

Getting an external point of view, from professional fields perhaps more indicated for the study of cinema, could complete the project bringing impressions and evaluations which can lead the analysis to a multidisciplinarity condivision of process and results.

Conclusion

These maps and visualizations are more than a picture of cinema's world and its inner dynamics, they are hints and source of interests which in the future could develop again, fixing the lacks of the project and showing even new aspects.

The need of feedbacks and opinions from domain experts, such as the possibility to explore new datasets make 'Borders' a starting point for further analysis.

At the same time the chance given by this project of analysing different aspects of huge archives related to millions of cinema's and television's productions shows tangible results and a macroscopic point of view of this world and society.

References

Ahmed, A., Batagelj, V., Fu, X., Hong, S., Merrick, D. and Mrvar, A. (2007) *Visualisation and analysis of the Internet movie database*. Paper Presented at the 6th International Asia-Pacific Symposium on IEEE.

Bastian, M., Heymann, S. and Jacomy, M. (2009) *Gephi: an open source software for exploring and manipulating networks.* Paper Presented the International AAAI Conference on Weblogs and Social Media, North America.

Bencivenga, A., Mattei, F.E.E., Chiarullo, L., Colangelo, D. and Percoco, A. (2013) *La formazione dell'immagine turistica della Basilicata e il ruolo del cinema*. Rivista di Scienze Regionali*,* 3 (6).

Caviglia, G. (2013) *The design of heuristic practices. Rethinking communication design in the digital humanities.* Unpublished Ph.D. Dissertation.

Cutting, J.E., Brunick, K.L., DeLong, J.E., Iricinschi, C. and Candan, A. (2011) *Quicker, faster, darker: Changes in Hollywood film over 75 years. I-Perception*, 2 (6), 569-576.

Goldfarb, D., Arends, M., Froschauer, J. and Merkl, D. (2013) *Art History on Wikipedia, a Macroscopic Observation*. Paper Presented at the Proceedings of the 3rd Annual ACM Web Science Conference, New York.

Herr, B.W., Ke, W., Hardy, E.F. and Börner, K. (2007) *Movies and Actors: Mapping the Internet Movie Database.* Paper Presented at the Proceedings of the 11th International Conference on Information Visualization Zurich.

Jacomy, M., Heymann, S., Venturini, T. and Bastian, M. (2011) *ForceAtlas2, A continuous graph layout algorithm for handy network visualization*. Medialab center of research.

Jessop, M. (2008) Digital visualization as a scholarly activity. *Literary and Linguistic Computing,* 23 (3), 281-293.

Jockers, M.L. (2012) *Computing and visualizing the 19th-century literary genome*. Paper Presented at the Proceedings of Digital Humanities Conference, Hamburg.

Kittur, A., Suh, B. and Chi, E.H. (2008) *Can you ever trust a wiki?: impacting perceived trustworthiness in wikipedia*. Paper Presented at the Proceedings of ACM conference on Computer supported cooperative.

Latour, B. (1996) *On actor-network theory. A few clarifications plus more than a few complications*. *Soziale welt,* 47 (4), 369-381.

Manovich, L. (2013) *Visualizing Vertov*. *Russian Journal of Communication*, 5 (1), 44-55.

Masud, L., Valsecchi, F., Ciuccarelli, P., Ricci, D. and Caviglia, G. (2010) *From data to knowledge-visualizations as transformation processes within the data-information-knowledge continuum*. Paper Presented at the Proceedings of the International Conference on Information Visualisation.

Morawetz, N., Hardy, J., Haslam, C. and Randle, K. (2007) Finance, Policy and Industrial Dynamics—The Rise of Co-productions in the Film Industry. *Industry and Innovation,* 14 (4), 421-443.

Moretti, F. (2005) *Graphs, maps, trees: abstract models for a literary history.* London: Verso Books.

Van Ham, F. and Perer, A. (2009) 'Search, Show Context, Expand on Demand': Supporting Large Graph Exploration with Degree-of-Interest.

Ieee Transactions On Visualization And Computer Graphics, 15 (6), 953-960.