Schriften des Instituts für Dokumentologie und Editorik — Band 11

Kodikologie und Paläographie im digitalen Zeitalter 4

Codicology and Palaeography in the Digital Age 4

herausgegeben von | edited by

Hannah Busch, Franz Fischer, Patrick Sahle

unter Mitarbeit von | in collaboration with

Bernhard Assmann, Philipp Hegel, Celia Krause

2017
BoD, Norderstedt

Bibliografische Information der Deutschen Nationalbibliothek:
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de/ abrufbar.

Digitale Parallelfassung der gedruckten Publikation zur Archivierung im Kölner Universitäts-Publikations-Server (KUPS). Stand 4. September 2017.

SPONSORED BY THE

Federal Ministry
of Education
and Research

Diese Publikation wurde im Rahmen des Projektes eCodicology (Förderkennzeichen 01UG1350A-C) mit Mitteln des Bundesministeriums für Bildung und Forschung (BMBF) gefördert.
Publication realised within the project eCodicology (funding code 01UG1350A-C) with financial resources of the German Federal Ministry of Research and Education (BMBF).

2017

Herstellung und Verlag: Books on Demand GmbH, Norderstedt
ISBN: 978-3-7448-3877-1
Einbandgestaltung: Julia Sorouri, basierend auf Vorarbeiten von Johanna Puhl und Katharina Weber; Coverbild nach einer Vorlage von Swati Chandna.
Satz: LuaTEX und Bernhard Assmann

Image and Text in Numbers: Layout Analysis for Hispanic and Spanish Modern Magazines

Nanette Rißler-Pipka

Abstract

Hispanic and Spanish modern magazines were long time a neglected field of research. Even if the magazines are regarded as a valuable source for information about contemporary cultural, social and political life for various disciplines like cultural, literary, media or social studies or linguistics. But stored in libraries as sensitive material and threated soon by decay most of the magazines were not accessible for researchers. Since digitization this has changed. The project *Revistas culturales 2.0* (University of Augsburg) tries to work on the digital collection of the IAI (Ibero-American Institute, Preußischer Kulturbesitz, Berlin) by annotating the magazine pages and analysing the metadata using digital tools. But as we are experienced in cultural, media and romance studies, the complex field of automatic document analysis stayed hidden for us, without the cooperation with experts in computer science (where in context of digitizing projects and OCR important research is already achieved).[1] Also, in the DH this field is not as advanced as for example text mining, therefore it proved difficult to find tools for quantitative analysis (for example relation of textual and image parts in the magazines). In cooperation with the project *eCodicology* and reusing their tool SWATI we found now a way how layout analysis for Hispanic and Spanish modern magazines might be done in future by measuring each page automatically. In consequence the paper can only present a concept of how useful the tool and quantitative analysis of the layout might be for analysing the Hispanic and Spanish modern magazines from the perspective of Humanities.

Zusammenfassung

Lateinamerikanische und spanische Kulturzeitschriften der Moderne konnten lange nicht für Forschungszwecke im Original herangezogen werden. Dabei sind sie als wertvolle Quellen als Zeitzeugen des kulturellen, sozialen und politischen Lebens anerkannt und könnten in verschiedenen Disziplinen wie Kultur-, Literatur-, Medien- oder Sozialwissenschaften sowie in der Linguistik genutzt werden. Jedoch waren sie als empfindliches und leicht vergängliches Material in den Bibliotheken unter Verschluss und auch aufgrund geographischer Entfernung für viele Forscher

[1] See also the Pattern Recognition & Image Analysis Research Lab of the University of Salford Manchester (PrimA).

Kodikologie und Paläographie im Digitalen Zeitalter 4 – Codicology and Palaeography in the Digital Age 4. Hrsg. Hannah Busch, Franz Fischer und Patrick Sahle, unter Mitarbeit von Bernhard Assmann, Philipp Hegel und Celia Krause. Schriften des Instituts für Dokumentologie und Editorik 11. Norderstedt: Books on Demand, 2017. 25–42.

unerreichbar. Seit der Digitalisierung hat sich dies geändert. Das Projekt *Revistas culturales 2.0* (Universität Augsburg) versucht nun mit der digitalen Sammlung des IAI (Ibero-Amerikanisches Institut, Preußischer Kulturbesitz, Berlin) zu arbeiten. Dazu werden Metadaten angereichert und vorhandene Metadaten mithilfe digitaler Tools analysiert. Als Geisteswissenschaftler musste uns bisher das Gebiet der automatischen Bildanalyse verschlossen bleiben, da dies vor allem auf technischer Seite in der Informatik (im Kontext von Digitalisierung und OCR) weiterentwickelt wurde.[2] Auch in den DH ist dieses Gebiet weit weniger bearbeitet worden als z.B. Text Mining und es erwies sich als schwierig, Tools für quantitative Analysen (um z.B. die Bild-Text-Relation in den Zeitschriften zu beziffern) zu finden. Nun konnten wir in Kooperation mit dem Projekt *eCodicology* und durch Nachnutzung ihres Tools SWATI ausprobieren, wie das Layout der lateinamerikanischen und spanischen Kulturzeitschriften der Moderne durch automatisches Vermessen jeder einzelnen Seite in Zukunft analysiert werden könnte. Der vorliegende Beitrag entwickelt aus einem ersten Experiment mit einer kleinen Anzahl von Zeitschriftenseiten ein theoretisches Konzept, um den Nutzen von quantitativen Methoden und im konkreten Fall vom angewendeten Tool aus Sicht der Geisteswissenschaften abzuschätzen.

1 Modern magazines as complex work of art

When libraries all over the world began to scan their treasures of thousands of pages of magazines, produced in the second half of 19th century until the second half of 20th century, many researchers (particularly in the Anglo-Saxon world) realized the potential that lies in the gained data. For the period of Modernity, the impact of cultural magazines on literary, artistic and social life is commonly known. The fast changing society required a medium that represented this acceleration in weekly or monthly issues. As well as the increasing linkage between fine arts, literature, photography, architecture, fashion, life-style, etc. was looking for a medium that incorporates the cultural change altogether. Furthermore, technical progress in printing industries and aesthetic ideas coming from the rising avant-garde culture produced an amazing number of new and partly ephemeral cultural magazines all over the western civilization (including Latin America). But to consider cultural magazines as an independent and complex work of art, which deserves to be analysed as a whole (i.e. each title of a magazine and the relation between magazines), has not been standard in literary, cultural, media or social studies for many years. Traditionally cultural magazines were used like a library of rare texts from known authors or as

[2] Ibid.

information reservoir for contemporary witness. The function of magazines was rather the one of an archive than of an original work of art worth to be looked at as a whole (Ehrlicher and Herzgsell 2016; Podewski 2016).

Now the digitization of numerous magazines initiated a change in their perception as complex cultural artefacts (Louis 2014; Pita González 2014; Maíz 2011; Stead and Védrine 2008). Equally the various changes in human perception in general which we can observe in Modernity are causally linked to the perception of magazines. People read the magazines in a public space and browse through them. They read them in fragments, are attracted by a picture or a headline and then interrupt their reading because of other distraction. For the cultural area of Latin America Raquel Macciuci draws the comparison between reading a magazine (or paper) and browsing the internet (Macciuci 2015, 219). Though, we have to take into consideration that contemporary readers and researchers reading and analysing the magazines today have a totally different kind of perception (Louis 2014, 33). In contrast to the contemporary reader we are able to have a look on all the issues of each magazine at once. This overview enables us to observe changes, particularly in the layout, but also regarding the staff and content of the magazine (the latter would be called metadata in the digital era). But to get an overview in a visual sense, we have to find a representation that allows us to look at many issues at once and to still recognize differences and changes in layout.

Before digitization, particularly in the Spanish speaking countries, most of the magazines were not accessible for researchers as they were stored in libraries in Latin America, Spain and elsewhere. Still, no researcher is able to read the massive number of pages produced in Spanish speaking magazines published in the period around 1850–1945. The research project *Revistas culturales 2.0* lists 585 titles of magazines, of which 224 titles are digitized, every title consists of 1–300 volumes with ca. 20–200 pages each. Directly accessible and annotatable via the website of our research project *Revistas culturales 2.0* are 23 titles of magazines, provided by our cooperation partner: Ibero-American Institute (IAI, Preußischer Kulturbesitz, Berlin). The small number of 23 titles still contains 477 single issues and about 23,000 pages. To read and analyse all of them is certainly possible, but time consuming and not very productive when the research question would be a comparison of layout, aesthetical and conceptual changes within and between the magazines.

For the image driven period of "modernism" the aesthetics are defined easily by layout: graphical elements, ornamental framings, printing types, etc. Therefore, the digital representation of the whole view of a magazine is important for projects like the *Modernist Journals Project* (MJP) or the *Modernist Magazines Project* (MMP) or the *Blue Mountain Project* (BMP). Nevertheless, none of these projects are able to do quantitative layout analysis for their document repositories.

2 How to analyse the layout of modern magazines using digital tools?

Like most of these Anglo-Saxon projects also the repository of the *Revistas culturales 2.0*-project consists of images, i.e. jpg-files. OCR is very complicated for the text written in columns and the mixture of text and image in the magazines. The *Blue Mountain Project* is the only one providing also the text of their rather small collection, but depending on the quality of the original, the OCR results might be poor.[3] For the few Spanish and Hispanic repositories of cultural magazines or periodicals providing also text gained by OCR, the plain text is not accessible and results for searching in the texts show poor OCR quality (Rißler-Pipka 2014, 60).

As layout and whole visual appearance of the modern magazines are very important, it is meaningful to analyse the layout of the magazines in their historic context – and not only to read the content. In the layout of the document the text is still quantitatively represented. The position and quantity of text is as important as the one of illustration, photograph, painting and other visual elements in the magazines. The problem is rather that digital research for document or image data is not in the same way advanced as text analysis (i.e. text mining, etc.). For users not experienced in computer science, it seems difficult to find a tool that is accessible and explained in a way other disciplines (like DH) except computer science are able to work with. In DH very celebrated, but also critically discussed tool ImagePlot by Lev Manovich and the Software Studies Initiative has certainly more potential than just plotting images into one, but it actually structures a great number of images on the basis of mostly two different features (represented in x- and y-axes). To be able to use the tool you need already measured and saved data behind the chosen features, for example average image saturation and average image brightness for each image in your collection (as in the Van Gogh example Manovich gives: 2015, 25–26). Manovich tested the tool also with the title pages of the *Times* by comparing the saturation of colour covers and brightness of black and white covers (Manovich and Douglas 2009). But for more complex data like the title pages of Hispanic modern magazines it comes already to its limits. Which kind of feature should be extracted to compare the magazines (or even only the title pages)? Brightness or colour saturation is not very useful here. When Manovich formulates questions for a collection of more than 10 million images like: "what are the subjects of these images" (Manovich 2015, 25), he is not going to answer them by using ImagePlot. For answering the question or even try to answer it, he uses primarily metadata (in the case of *On Broadway* he is not answering the question

[3] See the commentary in the Blue Mountain Archive (BMP): "Issues with poor quality paper, small print, mixed fonts, multiple column layouts, or damaged pages may have poor OCR accuracy. The searchable text and titles in this collection have been automatically generated using OCR software. They have not been manually reviewed or corrected."

Figure 1: Title pages of 18 magazines (1898–1931) with 245 issues from Latin America (IAI Collection)

by computing at all, but already the metadata provided by the source of the image (Instagram and social media statistics, Taxi statistics, Google Street View, etc.) gives information about the subjects of the images). In the context of document analysis as a field of computer science this approach is not even mentioned (Doermann and Tombre 2014).

Nevertheless, trying to look at the corpus in a bird's eye view (fig. 1), we can ask: What do we actually see, when looking at the miniatures of title pages? We are able to detect colour vs. black/white printing and we roughly detect images or bigger headlines vs. mostly text. If we would add more title pages in one image, we would see no details at all. In this case you need feature extraction and some computing before plotting them together. The difficulty still is, which are the features worth to extract and to compare? When Manovich claims that, "Computer can identify regions that have similar colour value and measure orientations of lines and properties of

texture in many parts of an image" (Manovich 2015, 22), he is not able to explain how it works. Lately Waltraud von Pippich proved the difficulties analysing paintings by using the computer. She points out that Manovich's analysis depends on pixel resolution and hardware. Plus, the comparison of mean values cannot represent the features of an image: "Zur Extraktion farbformaler Bildeigenschaften ist diese Methode der Medienkunst ungeeignet" (Pippich 2016).

For humanists the challenge is to see the image from a computer's perspective, that means without any semantics. For the computer text or image is just a different distribution of colour, brightness and other features. To find out, what could be the interesting features to be automatically extracted, we tried first a traditional interpretation of layout, keeping in mind the perspective of the computer.

2.1 Stepping back: Examples for traditional layout analysis

By zooming in the title pages (fig. 2) we could see changes in the development of some magazines, while others stuck to their layout for the whole period of their lifetime.

The Peruvian magazine *Amauta* has one of the most colourful and interesting layout in the collection (plus it is commonly regarded as one of the most important avant-garde magazine in Latin America). Initially repeating the typified Inca-head on the cover and then experimenting with various ideas of recognizable images and symbols until the last number in 1931 repeats again the initial Inca-head (appearance of the head 7 times over the years, for 31 issues in all). Other recurring elements are the rising sun (no. 10 and 30), the sowing Farmer-Inca (3 times: no. 12–14), the stairs (no. 19, 23, 28), the mixture of writing and mask in two different versions (no. 20 + 27 and 26 + 29) and finally the writing in combination with the initial Inca-head (no. 22, 25).

But working only with title pages can be heavily misleading. In the case of *Amauta* the fascinating aesthetics represented by the changing cover truly reflects the status of an avant-garde-magazine, but the inside of the magazine is not at all colourful or adventurous regarding the layout (fig. 2).

Typeface, paragraphs and the position of the images are not at all revolutionist or courageous, but rather traditional. The segmentation of the pages in text, image, headlines, paragraphs, even the decorative capital letter at the beginning of each article, poem, etc. can be described as conventional layout. Only the recurring ornaments or hieroglyphics of the Inca-culture are slightly unusual. Nevertheless, in the combination of the intriguing and consistent indigene aesthetics and the traditional layout it is recognized as a celebrated avant-garde magazine for Latin America. The obvious emphasis on the indigene roots of Peru fits to the fascination for indigene cultures in general in the international avant-garde movements. The editor Mariátegui brings his impressions of recent journeys to Europe into the concept of the magazine:

Figure 2: The title pages of *Amauta* (1926–1931)

"Yo vine de Europa con el propósito de fundar una revista" (Maríategui 1926, 1).[4] He tries to combine avant-garde-ideas and socialism as well as pro-Indio and European ideas. Taking into consideration the rich background information about Maríategui and his engagement in magazines (Beigel 2006; Manzoni 2004; Melgar Bao 2006), which can also be called metadata in the DH-sense, we have to evaluate the layout of *Amauta* in a different way.

The courage in layout and aesthetics of *Amauta* only shows when comparing it to other magazines of the same period, place and cultural context. The evaluation of the composition of image and text in the magazine should consider the contemporary standard.

[4] Translation: "Coming back from Europe, I had the idea of funding a magazine".

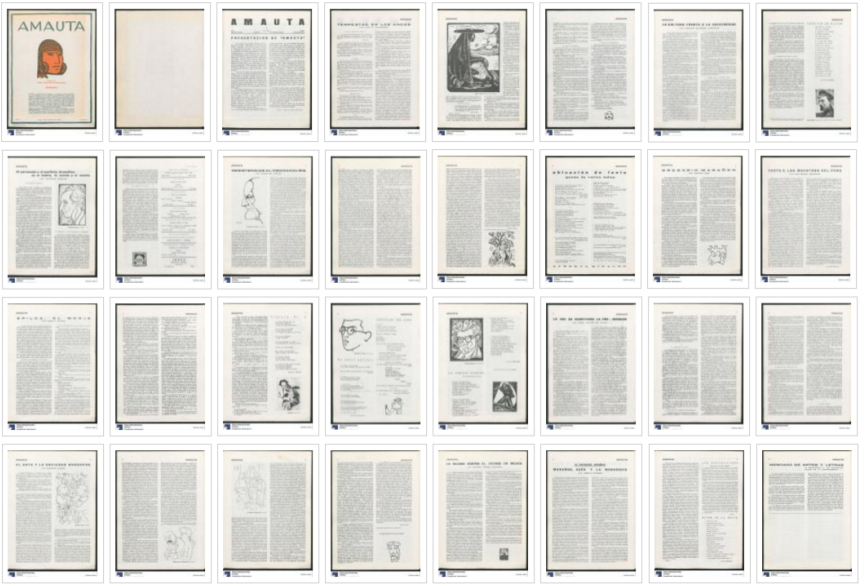

Figure 3: The first issue of *Amauta* (No. 1, 1926)

The rather conservative and elitist magazine *La Nueva Revista Peruana* shows in contrast to *Amauta* no images at all (fig. 4). This layout can be called book-look-alike and has certainly no interest to attract the reader's attention by visual aspects. Even if the editorial speaks of "una visión sin compromisos", this sentence is continued by looking back to the old values: "aunque transfigurada por el fuego de un antiguo fervor" (Editorial, Nueva Revista Peruana 1929, No. 1, p. 2).[5]

Even earlier, but representing another subgenre of cultural magazines, the *Ilustración Peruana* (1911–1912) is rather made to be browsed through and for distraction than for reading intellectual and cultural debates (fig. 5).

Yet here the layout can be misleading, because the distraction and rather decorative presentation of photographs, illustration and less text does not fit to the philosophical character of the first article "Nuestros Problemas – Valor y Trabajo" by Don Alejandro O. Escarza. Still conservative, the author requests some thinking from his reader by discussing on three long pages the problem of education in Peru. In contrast to Maríategui and his magazine *Amauta*, Escarza represents the conservative elite of Peru who are not willing to 'waste' money for the education of Indios.

[5] Translation: "a vision no-holds-barred … but expressed with the fire of former intensity".

Figure 4: *La Nueva Revista Peruana* (1929, No. 1)

These three Peruvian magazines show very clearly that neither the layout nor the metadata or textual content alone can be sufficient for an analysis. Plus, the variety concerning layout and concept of cultural magazines in such a short period of 1912–1929 in Peru is obviously quite large.

2.2 The other way – or why do we need digital tools?

What have we learned trying to analyse the three Peruvian examples of modern cultural magazines? We looked at 3 titles of magazines, part of 3 issues and at 72 pages and can say without looking at the rest of the pages of the three titles that each of them has a recognizable and individual layout. You could probably allocate correctly every single page of the magazines to the right title without reading a sentence of them. Though, allocating them does not mean knowing anything about them, apart from knowing that they belong together and that they are part of a magazine.

That means layout can be one criteria to distinguish different titles of magazines in different cultural contexts and different subgenres. But layout does not necessarily correspond to other semantics and can easily be misleading. To draw the deduction that more illustrations means necessarily a more popular magazine can be as false as

Figure 5: *Ilustración Peruana* (No. 152, 1912)

the deduction that a traditional layout only fits to a traditional magazine. Still, the whole visual appearance of a magazine is certainly part of its aesthetics and concept (in a cultural, social, political, artistic sense). If the content and other elements really fit to the visual signals is another question.

The variety in layout and other visual elements of modern magazines is rather numerous as we have seen in the three examples above. For an analysis of layout in a quantitative way the possibility to deduct general hypothesis on the basis of some exemplary analysis seems to be difficult and speculative. 'Counting' the layout as an interplay of text and image in numbers promises a more reliable method. The chance to analyse all of the ca. 23,000 pages in the IA-collection – and even more, if other repositories are used – is worth the effort of learning to handle digital tools and to ask the right questions to know which features of the document should be extracted and how to analyse the results. The example of Lev Manovich shows how difficult it is to find the right tool and that computer science is needed to understand the functionality and perspective of automatic document analysis. Therefore, we were happy to find in the cooperation with the project *eCodicology* not only the fitting tool, but also the explication of it (Chandna et al. 2016; Chandna et al. 2015). Even if the handling of

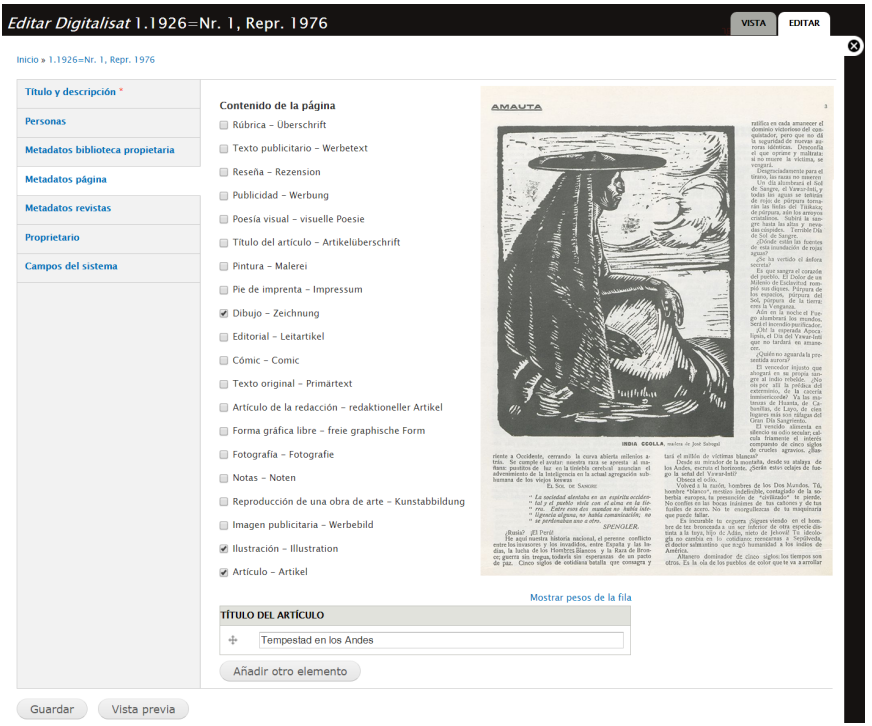

Figure 6: User interface for annotation, gathering metadata, example, *Amauta* (No. 3, 1926)

the tool had to be left to the experts (KIT, Karlsruhe Technology Institute) and we can only interpret the results, but not the technical function in detail. In this point further cooperation is needed.

Nevertheless, the manual or traditional layout analysis of the three Peruvian examples also showed, that we need more metadata to draw conclusions which are reliable and based on more information than just the visual one. On the project website of *Revistas culturales 2.0* we implemented the possibility to annotate each page of the collection (fig. 6).

While gathering metadata on the basis of crowd-sourcing the problem is that not enough people are willing to do the annotating of Hispanic modern magazines, because the community is not yet used to work with digital corpora. However, the metadata we already get by the IAI (title, year, contributors, place, country) and the one gathered manually by project collaborators help to do geographic overviews (see fig. 7) or network visualizations (Ehrlicher and Herzgsell 2016).

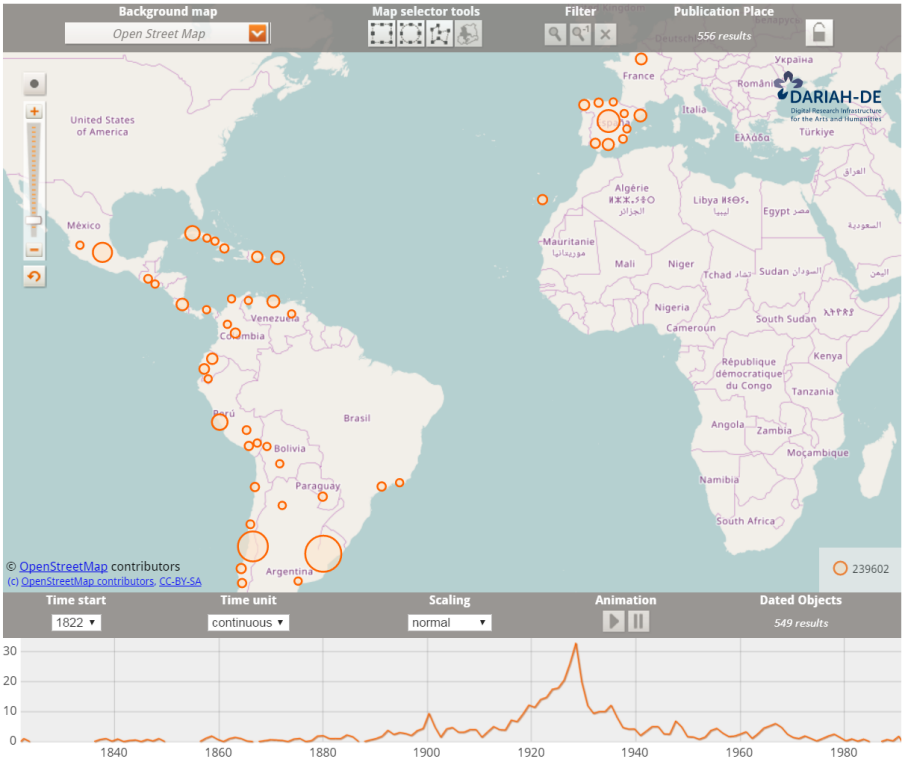

Figure 7: The bibliography of Hispanic and Spanish modern magazines in *Revistas culturales 2.0* visualized by DARIAH Geo-Browser

This example also shows why digital tools may help to analyse modern magazines in their cultural context. The distribution of magazines (as visible in the timeline, most of them appeared 1900–1940, with a peak in the 20s) and the spreading all over the Spanish-speaking world is amazing. The metadata is structured and ready for interactive use (for example to see all titles published in a chosen place or at a chosen time). Still, structured metadata alone does not help to know anything about the visual appearance of each magazine.

3 The experiment: Trying a tool for medieval manuscript analysis

The reuse of existing tools in a completely different context is rare, because every project designed their tool for exactly the data and research questions they may have. But we tried, if a tool designed for the feature extraction in medieval manuscripts can be adopted for measuring also pages of cultural magazines. From the computer science point of view this is a complex and difficult process, which is not transparent for me as a researcher in humanities. Nevertheless, knowing that it is possible enlarges the possibilities for cultural analysis. Together we think about the question: which features should be extracted? But as we already knew the features that 'can' be extracted by using the tool SWATI, we now have an idea of how it works:

"Layout features of the medieval manuscripts extracted by SWATI: Number of Pages, Mean Colour Value, Page Width, Page Height, Upper Left Corner Coordinates of Page, Relative Measurements of the Page, Text Width, Text Height, Text Areas, Upper Left Corner Coordinates of Text, Relative Measurements of the Text, Pictorial Width, Pictorial Height, Number of Pictorial Areas, Upper Left Corner Coordinates of Pictures, Relative Measurements of the Pictures" (Chandna et al. 2016, 3).

The principle differentiation we are interested in, is the one between text and image. In general, we know how important this difference is for analysing magazines – even if the quantitative distribution of image and text can be misleading as we have seen in the example of *Ilustración Peruana* (see fig. 5). By the named features SWATI is extracting we would also know how many text- and picture-areas we have on each page and additionally which dimensions they have and which position in the page. Certainly there are more complex questions to be asked regarding modern magazines like the graphic elements, ornaments and printing types, but for the beginning the tool would do more than expected.

So, we transferred some examples of magazine pages provided by the IAI and the metadata for the scanning process (colour checker, etc.) to the KIT, where Swati Chandna tried if the tool is working with this different kind of document. Thanks to her (and the whole *eCodicology* team) we got the measurements for image and text separately for six example pages of the magazine *El Hogar* (Dec. 1919). The results are visible in "Image" and "Numbers" (fig. 8-10).

Comparing the image and text segmentation in the examples we see how the tool is working. Particularly for Humanists the direct comparison between original and image/text segmentation is very useful to understand the technical and mathematical process behind. Even the quite complicated mixture of image and text in the advertisements is correctly recognized, if you regard the graphical elements in the headlines of the advertisements as image. Indeed, they function as both: image and text, because they should attract the reader's attention by visual appearance and give

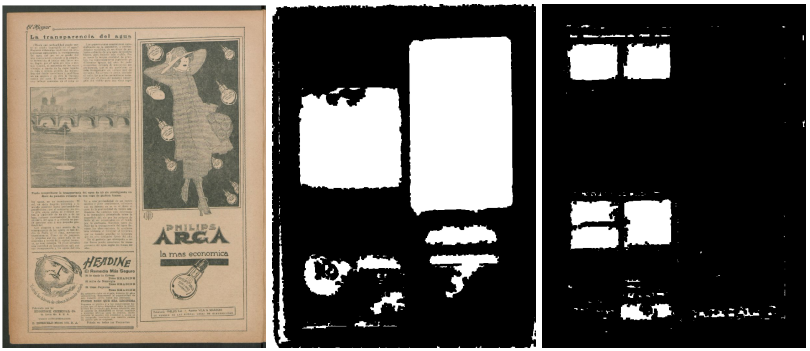

Figure 8: Original page – image segmentation – text segmentation

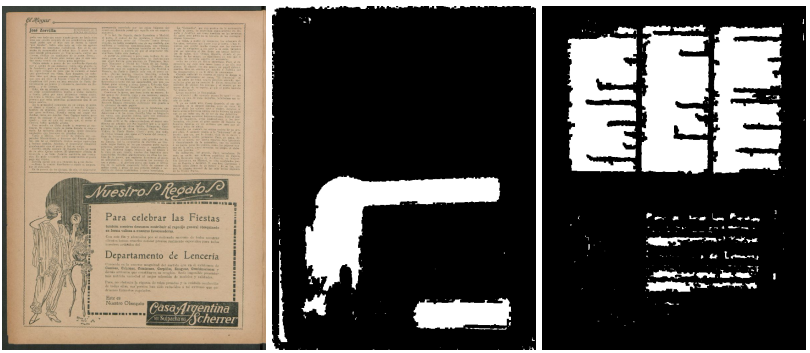

Figure 9: Original page – image segmentation – text segmentation

additionally textual information. The only problem, that might be resulting out of the measurement is the recognition of textual and pictorial units. For example, the advertisement in figure 9 is divided in a pictorial part and a textual part. In the tables of measurements, you won't know if these parts belong together or not – apart from using the extracted data for the position on the page and drawing the conclusion that textual and pictorial part belong together if they are positioned that near.

We hoped the results in numbers would be something like: For the title xy we count xxx images and xxx texts, so that we could compare these numbers with other results for other magazines, periods, places, editors, etc. But we had to learn, that measuring pages automatically is far more complicated. For both sides, the technical one and the humanist one, it is important to see by experimenting which limitation there are and to work together on further research to widen the possibilities.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Signature	Type	ID	Area	Mean	Min	Max	BX	BY	Width	Height	Major	Minor	Angle
2	6	page	1	885.241	255	255	255	0.517	0.475	26.314	33.873	38.046	29.625	89.963
3	6	picture	2	1.303	255	255	255	0	0	5.610	1.025	3.915	0.424	3.447
4	6	picture	3	83.635	255	255	255	3.356	17.280	19.831	11.763	16.629	6.404	22.026
5	6	picture	4	2.960	255	255	255	0.559	24.924	1.017	9.297	8.229	0.458	91.779
6	6	picture	5	3.200	255	255	255	2.975	29.254	2.551	2.254	2.268	1.797	12.263
7	6	picture	6	23.637	255	255	255	14.356	29.720	9.975	3.144	10.708	2.811	2.057
8	6	text	7	1.247	255	255	255	13.288	1.915	3.356	0.678	3.333	0.477	2.339
9	6	text	8	83.474	255	255	255	10.110	2.000	6.831	15.068	15.513	6.851	90.172
10	6	text	9	1.039	255	255	255	7.347	2.085	2.297	0.763	2.037	0.649	178.675
11	6	text	10	81.214	255	255	255	17.364	2.085	7.093	14.983	15.310	6.754	88.771
12	6	text	11	79.294	255	255	255	2.593	2.847	7.051	14.093	14.523	6.952	88.767
13	6	text	12	2.814	255	255	255	10.576	23.441	5.856	1.017	4.793	0.747	3.170
14	6	text	13	1.840	255	255	255	10.958	26.881	4.542	0.890	4.009	0.584	179.296
15	6	text	14	3.766	255	255	255	10.534	27.941	4.966	1.186	4.947	0.969	179.561

Figure 10: Table of measurements for textual and image parts, for one page – only biggest values considered (*El Hogar*, Dec. 1919)

For now, the measurements give us many numbers and even more images (see fig. 8-9, plus the page segmentation not illustrated here). That means on the one hand we reduced the complexity of each page, by focusing on pictorial and textual feature extraction, but on the other hand we produced much more complexity by all the data (in numbers and images) we gathered. All in all, the metadata gathered automatically is a result that cannot be valued highly enough. It is reliable because it is exactly the same process for each page, but should be observed and interpreted together with the people who programmed the tool.

What we tried here for some example pages should be easily done for all of the 23,000 pages of the IAI collection, but what we need now is a reduction of complexity for the gained data to be able to analyse it. That means we have to learn how to read and interpret the figures in the table (fig. 10). A first step to work with the given tables of measurement could be, to take only the biggest values for text and image, because all the tiny parts probably belong to one of the bigger ones. But as Swati Chandna points out, this won't work as a rule and it is necessary to observe the data first very carefully before defining the filtering rules.

Still, we don't know really how many entire images and texts are on the page. So, the next step would be to do statistical analysis and visualization of the metadata. As described for the visualization tool CodiVis in the *eCodicology* project this might be also possible for the modern magazines (Chandna et al. 2016, 3–5). Another perspective is, that the few examples show already that the distribution of textual and image parts on each page might give information of the kind of image or text, we look at (without really need to look at each page). The advertisements in *El Hogar* in all examples combine few textual parts (not forming an associated block) with image parts (also not forming the quadrat in which real illustration are usually represented).

Based on this observation we can try to build a model for recognizing advertisement semi-automatically. It might also happen, that advertisement for other titles than *El Hogar* use different forms of representation which can be described as different mathematic models to refine the tool with. Another challenge will be to recognize automatically the correlation between textual parts and image forming one ensemble. For example, in figure 8 we see that the image and the text below belong together and form one single advertisement, but in the automatic recognition of text and image parts this ensemble is correctly recognized as separate parts. The same is true for the example in figure 9.

The short experiment shows the different approaches in computer science (or at least DH) and Humanities. The measurements do not necessarily fit to the research question formulated in Humanities. That means more cooperation and exchange of ideas is necessary. But, looking at the results from a humanist point of view the effect is an estrangement (as Stephen Ramsay points out in *Algorithmic Criticism*). That means, we can now explore and observe aspects which were hidden for us before. For example, the fact that bold letters (in advertisements or other contexts) are recognized as image parts is meaningful also in aesthetics, because the function of these letters is at the same time a pictorial and writing one. The obvious failure of recognition becomes a double insight when observed in the humanist context.

4 Combined methods and knowledge

For a conclusion of the starting experiment, we can state that neither the traditional method of layout and magazine analysis can provide satisfying results for the whole corpus, nor the quantitative method only tested for now with some example pages can do the analysis of the magazines in their cultural context.

But a combination of the two methods could be a step towards an analysis of (Hispanic and Spanish) modern magazines considered as a complex interplay of text, image, content, political, social and cultural context. Problems to be solved in the future are:

1. Try the tool SWATI for the rest of the corpus, then try CodiVis and other statistics on the gained data.
2. Integrate the metadata of the magazines already gathered by annotating and the bibliography into the visualization for having more relation between the raw measurement and the context of each magazine. Therefore, we would be able to answer the difficult questions of an historic process: How much changed the appearance of cultural modern magazines from 1850–1945? What are the parameters of change?

3. Try a quantitative analysis and comparing this to the results gained already in secondary literature on Spanish and Hispanic modern magazines and in own exemplary analysis.

Acknowledgements

This work is the result of a cooperation between the project *Revistas-culturales 2.0* (University of Augsburg), the IAI (Ibero-American Institute, Preußischer Kulturbesitz, Berlin) and the project *eCodicology* (Technische Universität Darmstadt, Karlsruhe Institute of Technology, University of Trier). I thank the IAI for providing the data and the team of *eCodicology* for adapting and applying the tool SWATI on the data. For personal support and patient explication, I thank particularly Swati Chandna (KIT).

Bibliography

Beigel, Fernanda. *La epopeya de una generación y una revista: las redes editoriales de José Carlos Mariátegui en América Latina.* Buenos Aires: Editorial Biblos, 2006.

BMP: *Blue Mountain Project.* <http://bluemountain.princeton.edu>.

Chandna, Swati, Danah Tonne, Thomas Jejkal et al. "Software Workflow for the Automatic Tagging of Medieval Manuscript Images (SWATI)." In Ringger, Eric K., and Bart Lamiroy (eds.). *Document Recognition and Retrieval XXII.* 940206. 2015. DOI: 10.1117/12.2076124.

Chandna, Swati, Danah Tonne, Rainer Stotzka et al. "An Effective Visualization Technique for Determining Co-Relations in High-Dimensional Medieval Manuscripts Data." *Visualization and Data Analysis 2016*, San Francisco, California, USA, February 14-18, 2016. 1–6. <http://ist.publisher.ingentaconnect.com/contentone/ist/ei/2016/00002016/00000001/art00013>.

Doermann, David and Karl Tombre. *Handbook of Document Image Processing and Recognition.* London: Springer, 2014.

Ehrlicher, Hanno, and Teresa Herzgsell. "Zeitschriften Als Netzwerke Und Ihre Digitale Visualisierung. Grundlegende Methodologische Überlegungen Und Erste Anwendungsbeispiele." *Revistas Culturales 2.0*, 2016. <http://www.revistas-culturales.de/de/buchseite/hanno-ehrlicher-teresa-herzgsell-zeitschriften-als-netzwerke-und-ihre-digitale>.

Louis, Annick. "Las Revistas Literarias Como Objeto de Estudio." In Ehrlicher, Hanno and Nanette Rißler-Pipka (eds.). *Almacenes de un tiempo en fuga: Revistas culturales en la modernidad hispánica.* Aachen: Shaker, 2014. 31–57.

Macciuci, Raquel. "Técnica, soporte, ámbitos de sociabilidad y mecanismos de legitimación: sobre la construcción de espacios de literatura en la prensa periódica." In Schlünder, Susanne, and Raquel Macciuci (eds.). *Literatura y técnica: derivas ficcionales y materiales. Libros, escritores, textos, frente a la máquina y la ciencia. Actas del VII Congreso Orbis Tertius.* La Plata: Ediciones del lado de acá, 2015. 205-231.

Maíz, Claudio. "Las Re (D) Vistas Latinoamericanas Y Las Tramas Culturales: Redes de Difusión En El Romanticismo Y El Modernismo." *Cuadernos Del CILHA* 12.14 (2011). 75–91.

Manovich, Lev. "Data science and digital art history." *International Journal for Digital Art History* 1 (2015). 14-35. <https://journals.ub.uni-heidelberg.de/index.php/dah/article/view/21631>.

Manovich, Lev and Jeremy Douglas. "Timeline: 4535 Time magazine covers, 1923-2009." 2009. <https://www.flickr.com/photos/culturevis/3951496507/in/set-72157624959121129.

Manzoni, Celina. "Las Formas de Lo Nuevo En El Ensayo: Revista de Avance Y Amauta." *Revista Iberoamericana* 70.208 (2004). 735–747.

Maríategui, José Carlos. "Presentación de *Amauta*." *Amauta* 1 (1926). 1.

Melgar Bao, Ricardo. "Mariátegui Y La Revista Amauta En 1927: Redes, Accidentes Y Deslindes." *Revista de Antropología* 4 (2006).

MJP: *Modernist Journals Project.* <http://www.modjourn.org>.

MMP: *Modernist Magazines Project.* <http://modernistmagazines.com>.

Pippich, Waltraud von. "Rot Rechnen." *Zeitschrift Für Digitale Geisteswissenschaften.* Sonderband 1: Grenzen und Möglichkeiten der Digital Humanities. 2016. DOI: 10.17175/sb001_016.

Pita González, Alexandra. "Las Revistas Culturales Como Soportes Materiales, Prácticas Sociales Y Espacios de Sociabilidad." In Ehrlicher, Hanno and Nanette Rißler-Pipka (eds.). *Almacenes de un tiempo en fuga: Revistas culturales en la modernidad hispánica.* Aachen: Shaker, 2014. 227–45.

Podewski, Madleen. "Mediengesteuerte Wandlungsprozesse. Zum Verhältnis zwischen Text und Bild in illustrierten Zeitschriften der Jahrhundertmitte." In Mellmann, Katja and Jesko Reiling (eds.). *Vergessene Konstellationen literarischer Öffentlichkeit zwischen 1840 und 1885.* Berlin: de Gruyter, 2016. 61–79.

PrimA: *Pattern Recognition & Image Analysis Research Lab.* University of Salford Manchester. <http://www.primaresearch.org>.

Ramsay, Stephen. *Reading Machines. Toward an Algorithmic Criticism.* Urbana: University Press, 2011.

Rißler-Pipka, Nanette. "Sobre Los Problemas de Investigación Con Revistas Culturales Digitalizadas Del Mundo Hispanohablante." In Rißler-Pipka, Nanette and Hanno Ehrlicher (eds.). *Almacenes de un tiempo en fuga: Revistas culturales en la modernidad hispánica.* Aachen: Shaker, 2014. 59–80.

Stead, Évangélia and Hélène Védrine (eds.). *L'Europe des revues (1880 - 1920): estampes, photographies, illustrations. Histoire de l'imprimé.* Paris: Presses de l'Univ. Paris-Sorbonne, 2008.

Revistas culturales 2.0: *Virtuelle Forschungsumgebung zur Erforschung spanischsprachiger Kulturzeitschriften der Moderne.* Universität Augsburg. 2014-2016. <http://www.revistas-culturales.de>