ISSN 2186-7437

NII Shonan Meeting Report

No. 132

NII Shonan Meeting Report

Prof. Dr. Gerhard Heyer
Prof. Dr. Mitsuyuki Inaba
Jun.-Prof. Dr. Martin Roth

March 11–14, 2019

National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-Ku, Tokyo, Japan

Modelling Cultural Processes

Organizers:
Prof. Dr. Gerhard Heyer (Leipzig University)
Prof. Dr. Mitsuyuki Inaba (Ritsumeikan University)
Jun.-Prof. Dr. Martin Roth (Leipzig University)

march 11–14, 2019

When modeling real life dynamics, computer science has up till now mainly been concerned with business processes. However, with the intrusion of data-rich digital media and applications such as computer games and web communication in everbody's everyday life, the modeling of cultural processes becomes possible and requires an equally thorough analysis. The future of digitization will to a large extent depend on how successful computers and computing technologies can be adapted to different cultures of gathering, processing, altering and creating various types of information. While we assume that business processes can be described in a strictly logical manner focusing on preconditions and sequential actions, cultural processes such as the formation of public opinion in web fora, for example, happen to a high degree in parallel without much interdependencies, unpredictability of agent's decisions, and surprising feedback loops, just to name a few characteristics. In our workshop we have addressed this challenge by bringing together computer scientists and media scientists from Japan and Germany in a truly intercultural discussion. In order to better understand and model cultural processes as complex, self-regulating, media based communication dynamics, we have looked at digital manifestations of cultural processes in various media (video games, SNS, VR/AR), and discussed analyses that have already been carried out by participants. Different kinds of interpretations from computer science and media science, as well as the German and Japanese perspectives, allowed us to look for and discuss differences and complementary views among the participating scientists. This discussion helped summarize the state of art and highlight areas where further research is needed. On the the media and cultural studies side, we have focused on the following problems found in cultural analysis:

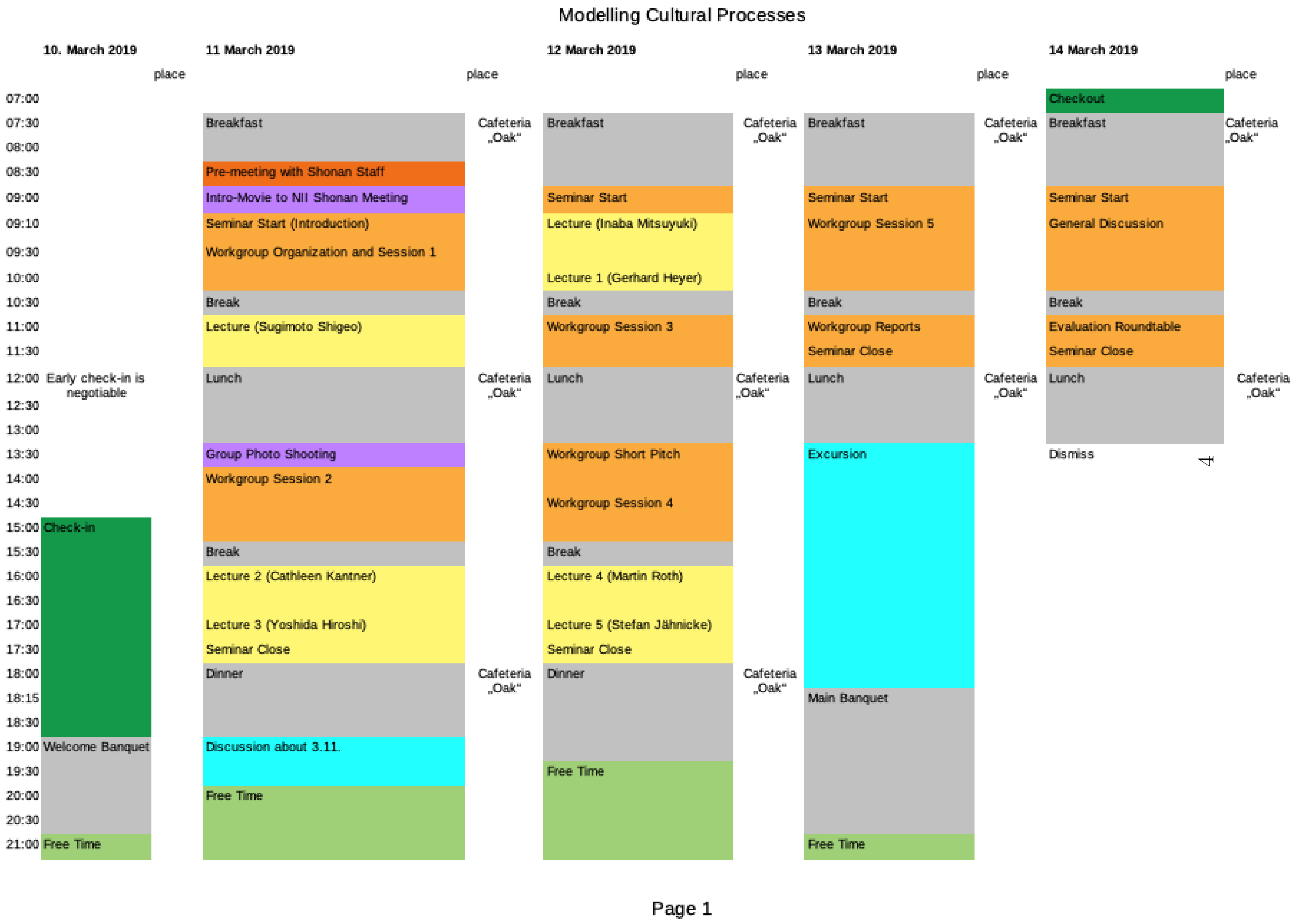
1. Workgroup 1: Identity in the digital age;
2. Workgroup 2: Shifting contents, shifting meanings across media, across time;
3. Workgroup 3: Constructions of culture;

For each of these areas, a team of researchers involving representatives from the humanities and computer science discussed and proposed a theoretical and methodological framework capable of modeling the cultural processes in question. In combination, the results achieved by each team allow us to develop a new perspective on the global significance and status of digital phenomena in their construction by actors and critics, as well as questions arising in comparison between different contents and cultures. In effect, our results contribute to the following problems, among others:

- *theory*, in particular a clear sense of the theoretical basis for our analysis;
- *data*, especially data sources, as well as access and analytical strategies to different kinds of data resources such as computer games and log files of how they have been played, or social media discussions on controversial topics;
- requirements, in particular the development of criteria that help to describe cultural dynamics in general as well as in detail, such as the nature and role of interpretations, or cultural identity;
- basic notions of modelling, such as the applicability of business process models and network theory, as well as, ideally, new models for specific contexts; and
- tools for modelling, such as different kinds of text mining tools, or visualizations for the dynamics of cultural processes.

The workshop addressed basic aspects of modeling cultural processes and developing software to support cultural dynamics. Apart from the enrichment of perspectives on our topic, the mixed German and Japanese group allowed us to discuss differences between the cultural engineering and computer-based modeling that have evolved due to different historical developments and cultural traditions of perception and action, in particular visualization. As a result, the participants have delivered multiple research agendas outlined below, which relate to representations, simulations, and evaluation. They may serve as a basis for intercultural and comparative media studies and a better development of future software that involves representations of cultural processes. Ultimately, the workshop challenged us to think about the status of hermeneutics and the relation between the humanities and computer sciences today, and highlights the ways in which both scientific perspective may benefit from and stimulate each other.

Schedule



Modelling Cultural Processes

	10. March 2019	place	11 March 2019	place	12 March 2019	place	13 March 2019	place	14 March 2019	place
07:00									Checkout	
07:30			Breakfast	Cafeteria „Oak“	Breakfast	Cafeteria „Oak“	Breakfast	Cafeteria „Oak“	Breakfast	Cafeteria „Oak“
08:00										
08:30			Pre-meeting with Shonan Staff							
09:00			Intro-Movie to NII Shonan Meeting		Seminar Start		Seminar Start		Seminar Start	
09:10			Seminar Start (Introduction)		Lecture (Inaba Mitsuyuki)		Workgroup Session 5		General Discussion	
09:30			Workgroup Organization and Session 1							
10:00					Lecture 1 (Gerhard Heyer)					
10:30			Break		Break		Break		Break	
11:00			Lecture (Sugimoto Shigeo)		Workgroup Session 3		Workgroup Reports		Evaluation Roundtable	
11:30							Seminar Close		Seminar Close	
12:00	Early check-in is negotiable		Lunch	Cafeteria „Oak“	Lunch	Cafeteria „Oak“	Lunch	Cafeteria „Oak“	Lunch	Cafeteria „Oak“
12:30										
13:00										
13:30			Group Photo Shooting		Workgroup Short Pitch		Excursion		Dismiss	
14:00			Workgroup Session 2							
14:30					Workgroup Session 4					
15:00	Check-in									
15:30			Break		Break					
16:00			Lecture 2 (Cathleen Kantner)		Lecture 4 (Martin Roth)					
16:30										
17:00			Lecture 3 (Yoshida Hiroshi)		Lecture 5 (Stefan Jähnicke)					
17:30			Seminar Close		Seminar Close					
18:00			Dinner	Cafeteria „Oak“	Dinner	Cafeteria „Oak“				
18:15							Main Banquet			
18:30										
19:00	Welcome Banquet		Discussion about 3.11.							
19:30					Free Time					
20:00			Free Time							
20:30										
21:00	Free Time						Free Time			

Page 1

Overview of Input-Presentations

Modelling Culture

Shigeo Sugimoto, University of Tsukuba

Between Scylla and Charybdis: Trade-Offs between the Need for Generic Tools and the Need for Hermeneutic Sensitivity in the Digital Humanities

Cathleen Kantner, University of Stuttgart

Metagaming and Ecologies of Video Games

Hiroshi Yoshida, Ritsumeikan University

Implementing Platforms of Cultural Construction

Mitsuyuki Inaba, Ritsumeikan University

Text Mining as a resource for modelling / NLP Toolbox in Action: Usecase: Overton Window

Gerhard Heyer and Christian Kahmann, Leipzig University

Exploring algorithmic/platform-centered online community cultures

Martin Roth and Peter Mühleder, Leipzig University / Leipzig University Library

Visualization and Digital Humanities

Stefan Jänicke, Leipzig University

Workgroup Outcomes

Workgroup 1: Modeling Cultural Processes for Collective Identities

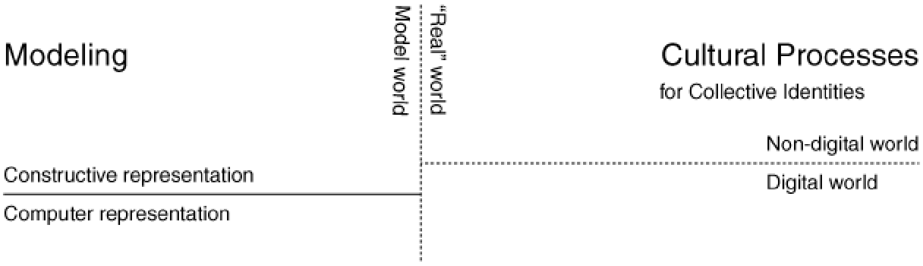
Akito Inoue, Peter Mühleder, Martin Potthast, Steffi Richter, Martin Roth, Gerik Scheuermann

Preface: Methodology of the work group

In our workgroup, we started with an exploratory discussion of the concept of identity and the associations all participants had with this concept and its various dimensions. Following this, we make first attempts at visualizing our concepts. The next step was to come up with a series of concrete research questions for interesting cases. Based on these questions, we discussed promising case studies and concrete approaches to them. This process resulted in singling out two interconnected cases, as well was in two general questions we would like to deal with in our project: a.) gaps in the concept of representation. b.) finding the discourses in the data.

I. Modeling and Representation

Any collaboration between computer scientists, humanities scholars and social scientists needs to be not only aware of the different perspectives and definitions of terminology, but should actively be problematized, constantly reflected on, and made productive in research. In our case, one particularly important gap between the fields concerns the concept of model and representation we apply. We organize the terminology in a broad way as follows:



In computer science, modeling it is an inadvertent requirement. In this regard, the term model has been inductively defined as follows:
"To an observer B, an object A is a model of A to the extent that B can use A* to answer questions that interest him about A."* (Marvin Minsky, 1965)
This pragmatic way of defining a model is strictly question-oriented: it presumes there are specific measures that will allow us to verify or check whether the answers produced by a model are actually valid with respect to the real world. Modeling is a way of purposefully abstracting over the real world systems and processes in question, reducing them to the least required, well-defined concepts that can be used to reason, both theoretically and practically. Although the definition is subjective (with respect to the vantage point of observer B),

the usefulness of a model transcends its usefulness to an individual in a given situation to answer a given question at hand, but is due to its generalizability to other people in other related situations with other related questions.

Since the goal of computer science is to solve problems with the help of a computer, modeling can be seen as a first step towards operationalizing a specific way of question answering: Can the model also be used by a computer to answer the question, rather than the observer. The model hence must be encoded in a computer, and that is what is typically called a computer representations (representation, for short). The representation, again, is an abstraction, and at the same time a discretization. Not only the model instance itself, but also its solution must be represented, and once the answer has been computer from the representation of a model instance through algorithmic means, it has to be transferred back into the real world by means of interpretation or action. Representations can take many forms; one that is currently often applied includes feature representations, where a feature is a function that maps real world objects onto a well-defined set of possible values to capture a specific characteristic or aspect.

For humanities scholars and many social scientists, representation and, more specifically, modeling, are political in the sense that they are not only supposed to approximate reality, but always also construct a specific reality. How I represent a phenomenon or structure also creates that structure or phenomenon in a distinct way. The positions of researchers range from moderate to more radical constructivist approaches, which regard reality as real only in the sense of its representation.

The key difference in understanding and modeling between the two disciplines is thus that computer science is somewhat pragmatically only concerned with how a certain part of the world can be encoded in a computer, thus differentiating "real" from model world and its representation, whereas humanities researchers try to understand the world in situ, being part of that same "real" world.

The challenge for the two fields is hence to reflect on the gap between these two concepts in all steps during the research process and consider how they can inform each other productively.

II. Operationalizing Collective Identity

1. Understanding Identity

Culture is inseparable intertwined with identity (and power) and so - whether we like it or not - Humanities has to deal with it. There are "what"-questions (what kind of identities/"snapshots") and (more important for us) "how"-questions, i.e. searching processes of (de-)construction of identities: process of drawing border between me/we and you/the other(s) (or even stranger(s)), of "in" and "out". Including somebody (and/ being included in a group/collective) takes places by articulating something (a topic) and somewhere, and means automatically excluding somebody. Articulation means linking/fixing language elements in a certain way - whereby "language" is understood in a broad sense: natural language, expert language, algorithms, body language (gestures). A central question is: WHO has the power to define, what the "right" way of fixing the elements is... (the question of hegemony/hegemonial identity (identities)

2. Where can we find expressions

The expression of our identity takes place at various levels. Some people express it as an action under consciousness. Some people express it as conscious verbal acts. Moreover, even we use ”word”, some words were expressed consciously and some words were expressed in under consciousness. They are mixed in it. The expressed identity affects the identity of others. They contribute to order the structure of both personal identity and collective identity.

3. Operationalization

Within a discourse, we are interested in identifying practices that reference (construct) collective identities, both of the self and the other, and/or express a position or relation to collective identities. Such practices may involve “rational” and affective elements, and different strategies, including for example irony. In order to identify the collective identities (WHAT) and trace the process of their construction over time (HOW), we need to identify instances in which collective identities are referenced, answer the question who is identifying whom as what how and why. We are specifically interested in measuring levels of solidarity and discrimination.

III. Research Questions

(1) (Political) collective identities in Yahoo News comments

(2) Formation of a ‘Gamergate’-style collective identity in Japan

The formation process of the Japanese gamer community is considered to be comparable to the English gamer community in several ways. Controversy about feminist discourses is constantly happening around otaku culture. By defining the ”enemy” in the process makes the structure of the argument. Who defines and who follows it? Today’s discussion doesn’t have an organization, but some of the people persistently take a similar role. Who could get the role gradually?

IV. Interlude: General approaches

The two research problems above lead to two distinctive general approaches on how a data-driven research process on collective identities can be formulated: one explorative (identifying collective identities in a data corpus), and one process orientated (understanding the formation of collective identities):

1. given a specific data corpus (e.g. Twitter, News comments, etc.), what collective identities emerge/surface within and how do they define/position themselves? How do they relate to or define other (specific or unspecific (the ‘other’)) collective identities?

2. being interested in a specific collective identity (‘the gamers’, ‘the gamergaters’, . . .), we want to understand the formation and development of this identity. How can we identify traces of this collective identity and build a data corpus accordingly?

V. Methodology

The research questions of the previous paragraph will be addressed by using computer based methods together with human intervention and interpretation. Several steps are necessary for this endeavor. In addition, the challenges of formulating the concept of identity require an iterative process as earlier decisions will most likely have to be changed based on insights gained by evaluation and interpreting the results later.

Data acquisition

The first step in the process is data acquisition. This requires several critical decisions on the data sources, followed by the actual retrieval of the data. For the Yahoo News example, the source platform is clearly defined. However, the data will be the news text, the comments on the news including user names, and at least also likes. One can enrich this with recommended articles from the platform if this is available. For the gamer gate question, the identification of sources is a major methodological challenge. In principle, all available data in social media, email threads, comments and likes of news media, etc. is of interest. The identification of the small relevant portion in this data flood is difficult and will require substantial research including close cooperation between computer science and social science. After the sources are identified, the data will be crawled and stored.

Regarding Gamergate in Japan, there is a series of medium-sized controversial topics in Japan. Not one big thing like Gamer Gate. Also, some of them do with a hashtag, but some of them didn't. There are several ways to obtain data. One way is to collect data from Twitter and To check who speaks (or RT, like) about those topics. We may visualize who speak those topic several times, and how many people depart/join the discussion. Probably, some people who join the discussion several times may have a strong position.

Analysis / Methods / Tools / Visualization

According to our current expectation, a larger part of the data processed will be textual data together with user names, relations like one text commenting another texts and rating data (e.g. likes). Therefore, natural language processing (NLP) and social network analysis will play a major role in the data processing pipeline. This holds specifically for aspects of NLP like sentiment analysis, and methods relation to discourse analysis. A critical aspect will be to find methods on how to derive information on collective identity formation within the observed (electronic images, i.e. user names) of real humans. [Missing: Research Challenges with respect to choosing methods, missing methods/tools] After the inference of collective identities with regard to the chosen computer representation, visualization will be used to inform social scientists about the outcome of the computer analysis of the acquired data. This requires an overview of the data, the possibility to zoom and filter, and retrieve details going back to original data pieces (e.g. set of textual comments with markers of detected sentiments) underlying the analysis results. [Adapted from Shneiderman's mantra - Overview first, Zoom and Filter, Details on Demand].

[Missing: expected visualization challenges]

Interpretation

"The purpose of computing is insight not numbers." [Richard Hamming] It should be clearly stated that this working group does strongly believe that computers cannot answer the original research questions. This requires human interpretation, including a careful consideration of the underlying assumptions, especially in terms of constructive and computer representation, as well as selecting of data (sources), and all choices along the processing pipeline. We embrace the use of interactive visualizations to allow effective human access to the data mass. The information communicated from the data via the visualization will be processed by the social scientists in an interactive fashion to create insight into the social phenomena of interest. We hope for a "discussion with the data" style of interaction that allows the social scientists to generate hypothesis from the data by exploration and to test hypothesis with the data. This requires care to avoid statistical traps.

Evaluation

- Critical review of results in terms of used methods/method parameters, data acquisition, data selection; especially checking statistical significance
- Critical review of results in terms of structural and computer representation of identity
- Checking the computer based reasoning by anectodal evidence based on data pieces selected from the data manually/other tools or selected by detail-on-demand possibilities of the visualization

Visualization for Communication

Besides carrying out the actual research, communication of the results in the respective social science community is of high importance. Traditional methods like scientific presentations, research papers, and potentially books will be used, but also digital publication including access to underlying data and used tools are possible. An important part of this communication is a representation of the data supporting the findings using visualization. Such a visualization may be different from visualizations used to generate the scientific results as the focus is on communicating a previously known aspect of the underlying data. While it is well-known in the visualization community that visualizations serve three different purposes, i.e. exploration, hypothesis testing, and communication, the communication aspect is usually less researched, making this aspect an interesting aspect of this research agenda. Due to the challenges related to the whole project, the final communicating visualizations can not be described in detail at this point, but these visualizations will probably involve overview on a large parts of the data of relevance to the communicated fact and selected individual data pieces supporting the claim (e.g. comments showing collective identity of some gamers in the gamer gate discussion). Another possibility would be a much more involved visualization based on the visual representations used in the actual research process that allows an informed judgment of other social scientists on the used constructive and computer representation of identity, the whole digital processing pipeline (i.e. provenance data, access to tools), and the interpretations.

Iterating the Pipeline
Experience shows that a challenging subject like collaborative identity and its manifestations on social platforms will not be studied in a sufficient way in a first attempt. Therefore, we expect an iterative process involving social scientists and computer scientists that will (hopefully) lead to answers to the research questions.

Workgroup 2: Shonan Model for Content Selection & Analysis

Gerhard Heyer, Kazufumi Fukuda, Fabian Schäfer, Cathleen Kantner, Maciej Piasecki, Chris Biemann, Hiroshi Yoshida, Stefan Jänicke

1. Introduction

In the NII Shonan meeting 132 on modelling cultural processes during March 10-14, 2019, a set of researchers discussed ways to better understand and model cultural processes as complex, self-regulating, media based communication dynamics. In order to foster the discussion, we set up three working groups each combining the different backgrounds and countries of participants to allow for an interdisciplinary and intercultural exchange about modeling cultural processes:

- (WG1) identity in the information age;
- (WG 2) shifting contents, shifting meanings across media and across time;
- (WG 3) constructions of culture.

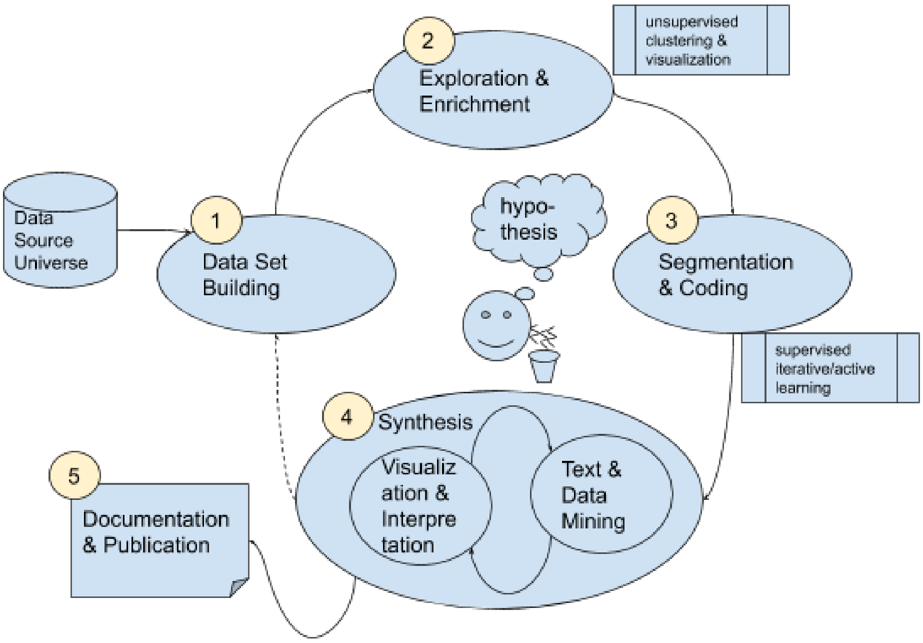
The background of discussion in our working group 2 was the common practice of modelling data and processes by using meta-data on the one hand, and standard text and data mining on the other. It was agreed that the goal of the discussion was to find a unified approach that combines both practices of modelling contents, and how they change over time. From this discussion, a model for content selection, enrichment and analysis emerged that subsumes a range of use cases from the social sciences and the humanities, supported by computer science with machine learning and visualization. This model assumes a model of research as a hermeneutic circle. It breaks down datafied research into steps so as to allow us to develop guidelines for operationalization and automatization. While following established setups of science, its value lies in the explication of steps and its recommendation on how to use automation within intellectually driven research processes, as well as its wide applicability in many fields, including the digital humanities, the social sciences and investigative journalism.
The researcher with her evolving notion of a research project is placed at the center of this model. She could conceive of the research project as a question(s), hypothesis or a social concept/phenomenon to be studied. We assume that her conception will evolve as she progresses. We also assume an iterative process where she may return to earlier steps and revise decisions based on the evolution

of the project. The model is represented as a research cycle of 5 steps, illustrated and presented below. This is a simplification for purposes of interdisciplinary collaboration in the development of workflows, tools, shared vocabulary and use cases.

After defining the model and detailing the steps, playing attention to the potential of automation, we will describe a range of use cases that can be subsumed under this model.

2. The Shonan Model

Now the parts of the process, which is also depicted in figure (1) below, will be detailed.


2
Exploration & Enrichment
unsupervised clustering & visualization
Data Source Universe
1
Data Set Building
hypo-thesis
3
Segmentation & Coding
supervised iterative/active learning
4
Synthesis
Visualiz ation & Interpre tation
Text & Data Mining
5
Documentation & Publication

To elaborate the concept and verify or falsify research hypotheses, the first step is the selection of the dataset (1) , which could consist of a corpus of text, media content or other primary data with associated metadata. Since any rough selection of source materials might produce unexpected results including unwanted by-catch, the next step is the Exploration and Enrichment (2) of the data. For this, we recommend the use of unsupervised clustering techniques coupled with a visualization to get an initial overview of the dataset and to quickly be able to remove outliers (e.g. with topic model visualizations or other fuzzy clustering techniques). This results in a refined, in-domain selection of materials, on which close reading and coding/annotation (3) is now performed to explicate the phenomena of interest with respect to the research question. This needs coffee. While the process of annotation/coding is manual and might involve the identification of passages and the assignment of hierarchically organized labels, it can be supported by supervised machine learning techniques that either propose labels based on previous assignments (iterative machine learning) or pick examples that might likely be assigned the same label (active learning). This step corresponds to what is known as coding in the social sciences and annotation in linguistics and can be understood as a process that adds research project specific metadata on data - be it on entire data items such as documents as well as on parts such as passages, sentences, words, picture regions etc. These annotations/codes/subjective metadata form the basis for the subsequent synthesis step (4), where findings are aggregated, visualized and interpreted. This step feeds into the documentation of the results as well as of the process towards a scholarly publication (5). Further, following the hermeneutic cycle paradigm, this re-shapes the research question/hypothesis/concept - a step done intellectually by the researcher, triggering another iteration of the entire process, starting with an altered data selection, different exploration and enrichment, augmented annotation/coding and a more advanced synthesis. Note that the four steps do not have to be executed in sequence – it is always possible to revise previous steps on the basis of such intellectual insight.

2.1 Data Set Building (Corpus Gathering)

- Researcher starts with initial intuitions or a vague idea of the concepts or phenomena to be investigated. She gathers a collection of data with which to study the phenomena. This may involve scraping, or selection from larger collections, crawling the internet, or even creation of data. Generally, there is a universe of data, where the dataset is selected from on the basis of search queries or metadata ranges (bulk selection, not individual selection, to be able to scale to large datasets).

- We assume there is a theory driving the framing of the research question and hence building of the data set. This might be a naïve theory that will be reformulated in the research (which may then lead to a rebuilding/altering of the data set.)

2.2. Exploration and Metadata Enrichment

The computational task at this stage is unsupervised clustering. Topic modeling as one such method can be used to define subsets relevant to a research community. These subsets should be public and contain metadata commonly agreed upon. Metadata for such subcollections can be retrieved by search engines and aggregated by metadata harvesting.

- Researcher explores the data to understand it and begin to formulate hypotheses that can be operationalized. Exploring the dataset is also a way of exploring her own naïve formulation of the project. She is focused on better specification of the concepts/phenomena and the ways in which they can be recognised in the data.

  - Manual methods - close reading - browsing and looking into the data. Also, surveying the metadata. For example she might skim the titles, authors, and dates of the items gathered by some process or she might randomly read some items closely.

  - Automated methods, e.g. unsupervised clustering, possibly interactive and iterative, coupled with visualisation methods, for a targeted and quick bulk selection of relevant material.

  - Visualization aids at illustrating characteristic features of the data set. Those might be statistical overviews like (interactive) charts and contentual overviews like tag clouds. The purpose is to deliver the shape of the current data set in a visual form and to guide the scholar to feature dimensions with lacking or over-represented information. Visualizations can further be designed to support annotation.

- Enrichment of the data items (of any granularity) with metadata. She may be drawing on disciplinary research to add metadata to the items that will allow her to ask certain questions.
  - acquired from the original sources,
  - manually created following the comparison of concepts/phenomena against the collected data
- Note that this is where one type of shift in content takes place as metadata gets layered on content, explicating aspects relevant to the current research question, which might or might not be shared with other researchers as additional metadata.

2.3. Segmentation and Coding

The computational task at this stage is iterative supervised learning/classification. Specific concepts (as defined by the researcher) are being applied by way of annotation to a fine grained layer of text (such as words or sentences). Such classes can be considered to be metadata as well. The generation of annotated text can be supported by machine learning such as active learning or bootstrapping approaches. - It needs to be noted that the researcher can always interact and correct the annotations

- Now the researcher begins to analyze (in the sense of break into parts) and then code or annotate things in the data set that are of research interest.
- She makes decisions about the granularity of data items to be coded (annotated) by concepts significant to the phenomena under investigation.
- Preparation of the coding (annotation) guidelines
  - iterative process with several experimental phases aimed at achieving good inter-annotator agreement,
  - possibly supported by automatic tools, e.g. in a form of test training and annotation, Active Learning, Bootstrapping, Iterative supervised learning etc.
  - Semi-automatic annotation editors: passage identification and classification / coded annotation. Iterative annotation: suggest codes and spans during the process of linearly annotating a set of documents. Active learning annotation: use tools that suggests similar passages in automatically chosen order.

2.4. Synthesis

The goal of this step is to find patterns in the annotated data. It can be supported by a number of approaches and tools, such as statistical analyses, neural nets, or pattern based analyses. Examples are network analysis, time series analysis, anomaly detection. Many of those tools are linked to specific visualizations and call for interactive fine tuning. Results of this stage may also generate metadata that can be used for further annotation of text. Results of the text and data mining stage need to be interpreted by the researcher. Visualization at this stage summarises data generated by the text and data mining stage.

- At this point, the researcher begins to synthesize an interpretation of the phenomena as represented by the annotated data. She will be using tools that propose views on the annotated data that may prompt insights into the phenomenon.
- Coded (annotated) data from the previous step are the basis for constructing (setting up or tuning the)
- Distant reading vs quantitative analysis vs computational analysis, e.g. by the Text and Data Mining methods
- Human analysis: Interpretation & Visualization
- Visualizations are designed to support hypothesis verification and generation. Distant reading visualizations show summaries of text/data mining results. Close readings allow inspection of individual data/text elements. A seamless transition between both perspectives is necessary to build trust in the underlying computational analysis methods: it is imperative that synthesized results allow accessing the underlying data to ensure full provenance.

2.5. Documentation

During the whole workflow, metadata have been generated semi-automatically for (1) the definition of subcorpora, (2) the fine-grained annotation of text, and (3) the results of the text and data mining analysis. All of these metadata can be used to evaluate the research workflow and adjust it to better fit the research question where necessary.

- Finally, the researcher wants to document the insights (interpretations) in ways that can be shared with others. She may be exporting passages (with citations), statistical results, visualizations, and so on for use in publications.
- She may also want to prepare a study dataset to be published with documentation for others to replicate her findings or for teaching.
- The enriched dataset can become a dataset gathered for another project thus closing the loop.
- She may want to document the processes of the previous research steps so as to try the same research flow on a different dataset.

It must be emphasized again that this process is not strictly sequential and does not follow a waterfall model. Rather, the researcher can always use intellectual revising to enter another loop, or to go back to previous steps in the current loop, in order to e.g. adjust the data set selection, revise the enrichment, alter the coding scheme etc. This is symbolized by the dashed line between steps 4 and 1 in the figure above.

3. Use Cases

Use Case 1: Humanities

1. Corpus Creation
   *Hathi Trust to study irony in Tudor drama

2. Exploration and enrichment
   Adding metadata about the plays from my knowledge of Tudor drama

3. Segmentation and Coding
   Searching for known ironic passages in known plays and coding them. Using tools to find similar passages. Training a machine to propose ironic passages. Eventually it might be possible for the machine to code the rest.

4. a. Text and Data Mining
   ...
   b. Visualization, Interpretation, Quotation
   Visualize distribution of passages over time, over authors, and over comedies/tragedies. Begin to form interpretation about Tudor irony.

5. Documentation
   Export the visualizations that make my point. Gather statistics and example quotes. Export a process visualization for the book. Export an "archive" that will be published online with the paper/book for others to use in recapitulating my results.

6. Feedback Loop to research question, theory

Use Case 2: Social Science

1. Corpus Creation
   Starting with a research question (rooted in a certain metatheoretical paradigm and an ongoing scientific controversy in our discipline)

   - Select relevant data sources in order to trace down the "social fact" you are interested in like "identity" or "power" (e.g. electronic text archives, . . .)
   - Segmentation of texts, importing text in database, metadata, preprocessing
   - Cleaning the raw corpus (doublets, sampling errors / false positives)

2. Exploration and enrichment
   Getting an overview of the corpus (e.g. frequencies, collocations, topic models, first time series, . . .) Adding notes or tags (new meta data) about those preliminary first findings

3. Segmentation and Coding
   Searching for highly interesting

   - Subcorpora
   - Text passages

Manual coding / annotation of a (random or layered random) sample of texts

- Coding interface would be helpful
- ti-marking function would be helpful (also in order to link it to learning algorithms in the future)
- Functionality for continuous intercoder-reliability (on different levels) would be helpful (compare e-Identity, RECON)
- For us also output functions for smaller subcorpora were helpful in order to process them further in commercial software tools like Atlas.ti or Provalis (not everything can / needs to be reprogrammed in the project specific tools)

4. a. Text and Data Mining
   Yes
   b. Visualization, Interpretation, Quotation
   Yes.

   - Nice visualization helps, however, certain fancy functions cannot be printed in scientific articles or books (obstacles: moving depictions, 3D, copyright issues for pictures . . .)
   - Data output functions are helpful in order to allow for the statistical analysis of the data generated from the text data in common statistical analysis (e.g. descriptive statistics, regressions, ARIMA time series, . . .) output not just of the graphs but also the relevant coefficients, tests of the preconditions that allow to use a certain statistics . . . .(necessary for documentation of the whole)
   - Partly statistical analysis gains value from combination with other data sources (surveys, event data, demographic or macro-economic)

5. Documentation

   - Export visualizations
   - Find the respective typical and / or exceptional quotations
   - Process visualisation is helpful (however, the whole research process / cycle usually is not just in one tool)
   - If done with SPSS or R in highly quantitative studies it becomes more an more common to publish also the syntax
   - Publication in scientific journals, papers books nice, if for both project partners (IT, SocSc) it results in innovative contributions to their resp. fields (here it is also good to have agreed in advance on a common publication strategy: e.g. in which types of journals which authors will be named first, . . . how to cite each others earlier work. . .)
   - Of course, we write our social science text ourselfs :-), however for interdisciplinary projects it is important to have time for that. Software creation, data work etc. cannot last up to the end of the project duration. There must be in between results to work with.

6. Feedback Loop to research question, theory
   - Wrap up, what worked and what not
   - Usually a possible reframing of theory, research questions, or consideration of new corpora is not done within one project, but rather in the planning of the next projects building on recently made experiences

Use Case 3: Disaster Archive

1. Corpus Creation
   Collecting photographs and other various resources for the future use by local governments, schools and research institutions

2. Exploration and enrichment
   Analyze the type of collected materials collecting know-hows for metadata creation Create metadata for each collected material Subject headings, thesauri don't exist. Folksonomy? Need resources such as geographical names and their changes over time

3. Segmentation and Coding
   Keywords, location, temporal information analysis Photographic image analysis — seems difficult without contextual information

4. a. Text and Data Mining
   Yes Technologies required: Image analysis Topic detection Metadata aggregation
   b. Visualization, Interpretation, Quotation
   Metadata visualization to help access

5. Documentation
   Annotations for linking the resources with community memory Geographic/temporal annotations...

6. Feedback Loop to research question, theory

Use Case 4: Investigative Journalism (new/s/leak project) www.newsleak.io

1. Corpus Creation
   Background corpus given by a set of leaked documents. Subcorpora are selected by metadata and fulltext search.

2. Exploration and enrichment
   Tool shows network of named entities and keywords from sub-corpus. Users can iteratively refine sub-corpus by including/excluding selectors.

3. Segmentation and Coding
   Documents can be tagged, thus marked as belonging to a specific case under investigation. Entity labels can be altered, added, allowing for the annotation of relevant keywords.

4. a. Text and Data Mining
   The entire tool is an interactive data mining environment
   b. Visualization, Interpretation, Quotation
   .. which has an interactive visualization.

5. Documentation
   Views and selectors can be saved for later use. In investigative journalism, leaks are the sources for further investigations that typically happen outside of the data collection, so underpinning of results is not suo much of an issue in this use case.

6. Feedback Loop to research question, theory
   This use-case integrates the feedback loop very tightly. A larger instantiation of feedback loop would be to add more background data on the basis of findings, but this does not match the reality of document leaks, which are typically one-time events.

Use Case 5: Creating Subject Data for Video Games

1. Corpus Creation
   Gathering text and/or metadata from the source of the records. (e.g. Wikipedia, Wikidata, Mobygames or any type of references)

2. Exploration and enrichment
   Getting an overview of the data, Analysing the type of the components (chapter).

3. Segmentation and Coding
   Making structured data from the analysis.

4. a. Text and Data Mining
   Find the keywords (subjects/topics) for works of video game through the automatization (e.g. text mining, topic model)
   b. Visualization, Interpretation, Quotation

   - Interpretation of the result from the mining is needed for the creating useful data.
   - Some type of visualization (e.g. network or cluster) will support to the analysis.

5. Documentation

   - Publish the created data on the online catalogue
   - Write the text of guideline for explain the spec of the describing elements (items)
   - Published the research paper of this analysis (if we can)

6. Feedback Loop to research question, theory
   Collect user's response on online catalog. Evaluate/Critique published data for generating new data or research.

Seems to be another instantiation: Text Mining for Qualitative Data Analysis CK: Gregor Wiedemann worked with Gerhard Heyer in Leipzig in the eHumanities Project "Postdemokratie und Neoliberalismus", constructing the Leipzig Corpus Miner

Workgroup 3: Tracing constructions of culture through League of Legends

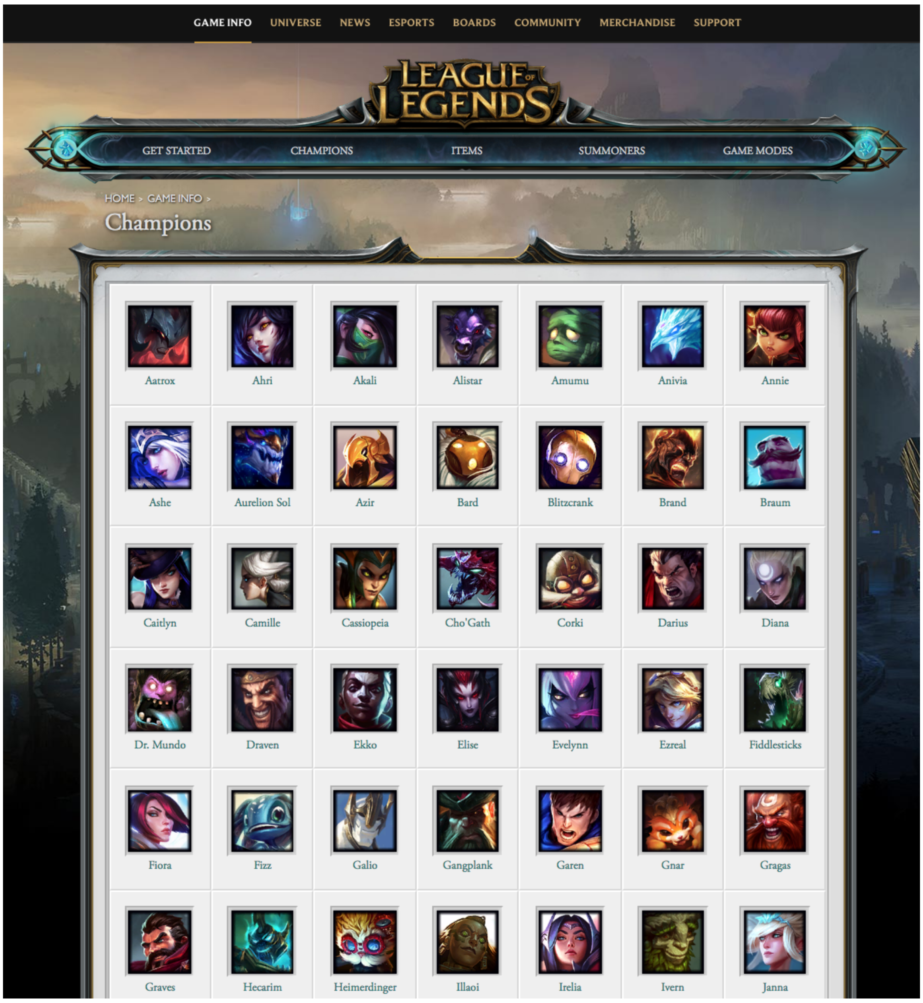
Mitsuyuki Inaba, Peter Chan, Verena Klemm, Thomas Efer, Christian Kahmann

Principal reflections on preconditions for working on and with culture/s: Ethics of research

Working on and with culture/s has strong ethical implications that we have to reflect at the beginning and during the work-process. The following questions are crucial:
1. Whose culture do we construct? How and why am I related (e.g. involved, separated etc) to this culture?
2. What is my own notion of culture? There is a multitude of definitions, some of them ahistorical and "culturalist", i.e. fixing people inside their culture and community (ethnicity, religion, traditions, nation state), not regarding the individual and the possibility of change and development in the course of historical/political processes. We have to avoid "Othering" (classifying human beings with other characteristics as different, strange) in comparison to one owns person/group.
3. Are hierarchies, power relations (between actor and his object) involved? (e.g. postcolonial, gender, west-/east etc.) 4. Which motivation is behind my research? Which interests and actors are involved (research interest, economical interests)?
How can we, in our case study, put these principles into effect?

Use Case: Creation of a set of rules inside League of Legends

Background: Introducing the Game
League of Legends (LoL) is multiplayer online battle arena (MOBA) video game. The game is played with 2 teams competing versus each other. Each teams is consisting of 5 players. The goal (win condition) of the game, is destroying the enemy team's base. Before every game, which takes between 20 to 60 minutes, each player has to choose a certain character (champion) from a pool of over 100 different champions, with unique look and abilities. During the game the players of a team need to communicate their strategy via Chat.



Instance of Champions

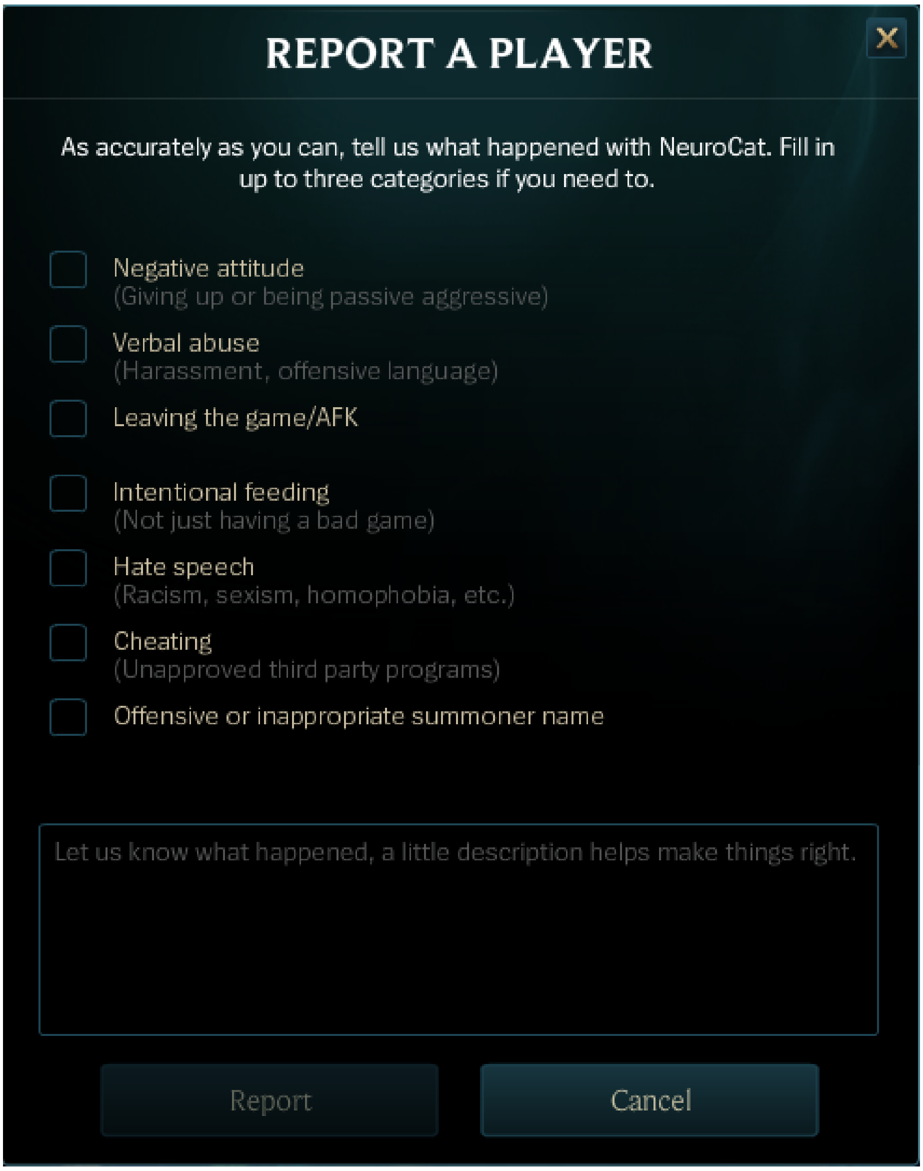
Impact of the game

Video gaming is an increasingly popular activity in contemporary society, especially among young people. By July 2012, League of Legends was the most played PC game in North America and Europe in terms of the number of hours played. In January 2014, over 67 million people played League of Legends per month. The 2017 World Championship had 60 million unique viewers and a total prize pool of over 4 million USD. The revenue of the game was: 2018: 1.4; 2017: 2.1 (billion USD). LoL has found impact in fan fictions, Cosplay and many other areas.

Research Questions

1. How does the creation of a set of rules and punishments influence the behaviour of players?

LoL is a game based on team work. Therefore it is necessary that all members of a team work jointly together. But very often this is not the case. People harass other players, blame them, insult them or just play bad on purpose, which is called “being toxic”, in order to make their own team lose. It’s called toxic, because an act of of misbehaving seemed to be an infectious action, spreading towards other players, who were more likely to also misbehave in one of the next games after having played together with a toxic player. So there was a need for establishing a set of rules and punishments for misbehaving. This construct of rules was introduced incrementally. After each game every player was allow to report another player for misbehaviour, specifying at least one the forbidden things he did. After that, a tribunal known "good behaving" players had to judge in an anonymised interface on whether and if so how to punish a reported player.



REPORT A PLAYER

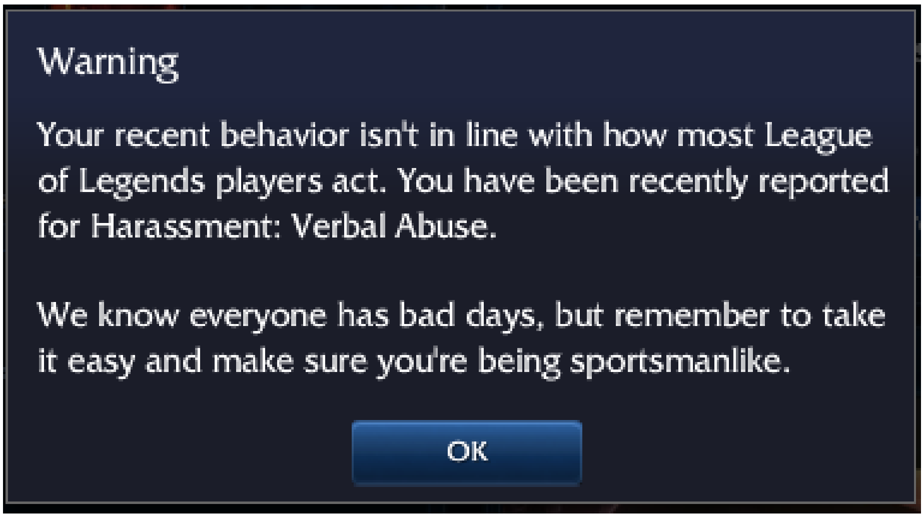
As accurately as you can, tell us what happened with NeuroCat. Fill in up to three categories if you need to.

- □ Negative attitude (Giving up or being passive aggressive)
- □ Verbal abuse (Harassment, offensive language)
- □ Leaving the game/AFK
- □ Intentional feeding (Not just having a bad game)
- □ Hate speech (Racism, sexism, homophobia, etc.)
- □ Cheating (Unapproved third party programs)
- □ Offensive or inappropriate summoner name

Let us know what happened, a little description helps make things right.

Interface for reporting a player for misbehaving

We are interested in measuring how players behaviour was influenced over time by the establishment of the set of rules? Was it possible to re-educate players or did the rules just “filter” the toxic players? What are the dynamics of spreading misbehaviour from one player to another?



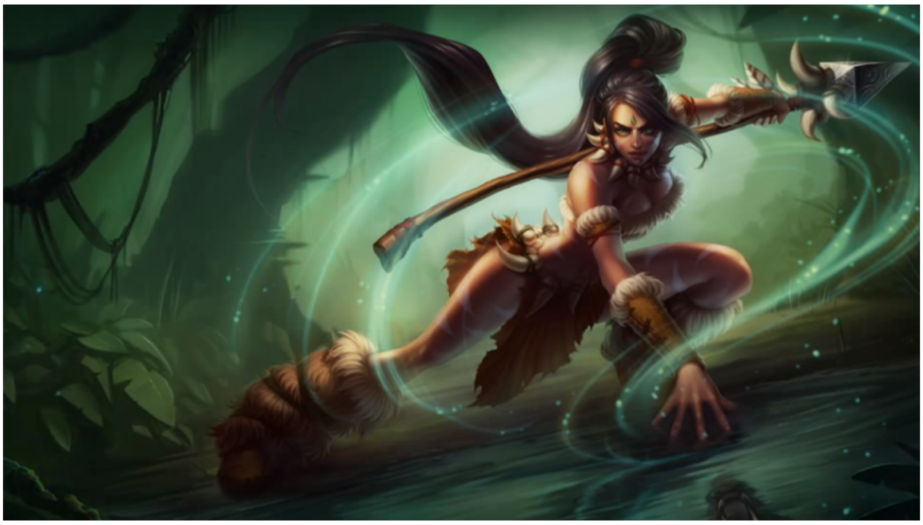
Warning

Your recent behavior isn't in line with how most League of Legends players act. You have been recently reported for Harassment: Verbal Abuse.

We know everyone has bad days, but remember to take it easy and make sure you're being sportsmanlike.

Message shown to a player who infringed the set of rules

2. (How) does gaming affect your behaviour in everyday life?
Is there an influence of gaming on your everyday life? Does it make players more aggressive, more relaxed, more communicative...?!
3. (How) does the image of male and female characters inside the game affect the perception of genders in the real world?
The characters played in the game are created on a very stereotype basis. A male character is created to be very masculine most of the time, having lots of muscles, carrying heavy weapons, smoking etc. . Whereas a female character is very attractive, showing lots of skin.



Female chamption:Nidalee

Male champion: Graves

4. (How) does culture mentality affect dynamics between players/ how to play? Does culture affect the way of playing?
Is there a correlation between measurable metadata of players like: Nationality, age , sex, experience in playing. . . and the ingame behaviour. Or is culture completely irrelevant in gaming? Why is it then the case that on competitive level south korean team win most of the time? Is there a culture conflict? Do people with a certain cultural background have a different way of playing (aggressive, defensive, relaxed, kind..). We need to keep in mind though, that culture does not cause behaviour. One is not forced to behave in a certain way, just because he has a particular nationality.

Further research ideas

- What are the incentives of people to play
- How to combine group of Champions? =¿ Balance behaviour
- What makes players pay just for the look of virtual avatar? =¿ Happy community will pay
- coop and co gaming to develop social skills? =¿ Games forcing families/people to communicate/ go out?!

What methods appropriate?

We need to know: How to measure behaviour? What are the rules? What are punishments? Do punishments affect ingame behaviour?

- Data and correlation analysis
- Factor analysis
- Bayesian statistics

- machine learning and text mining to measure behaviour
- Close reading on chat logs (edited videos from competitions)

What data is needed?

- Chat Logs
- ingame commands + game replay
- Data from Forums, magazines, twitch livestreams
- interviews, questionnaires

Feasibility?
Which methods are already available?

- Advanced Machine Learning (e.g. Neural Networks) to identify rule-breaking configurations in replay data (3D situational data) and textual log files (via Text Mining)
- Interviews and methods from the social sciences

Which tools resources and libraries are there?

- Twitch stream / youtube replays (collect resources for audience reactions)
- In-Game replay files can be obtained (a cooperation with the game studio would be desirable)
- API access and pre-existing datasets
  - API Explanation:
    `http://tercenya.github.io/compendium/introduction.html`
  - Static File API "Data Dragon" (e.g. user-created fan images):
    `https://developer.riotgames.com/static-data.html`
  - Tool for duplication of Static File Contents:
    `https://github.com/emreloper/serverless-datadragon`
  - Historical Matches dataset:
    `https://www.kaggle.com/chuckephron/leagueoflegends`
- On the financial success and business impact:
  `https://www.crunchbase.com/organization/riot-games`
- Sandbox Game Server:
  `https://github.com/LeagueSandbox/GameServer`
- Game outcome analysis in R:
  `https://www.kaggle.com/jaytegge/league-of-legends-data-analysis/notebook`

List of Participants

- Prof. Gerhard Heyer, Leipzig University
- Prof. Mitsuyuki Inaba, Ritsumeikan University
- Prof. Martin Roth, Leipzig University
- Dr. Thomas Efer, Leipzig University
- Prof. Akito Inoue, Ritsumeikan University
- Dr. Kazufumi Fukuda, Ritsumeikan University
- Prof. Steffi Richter, Leipzig University
- Prof. Fabian Schäfer, University of Erlangen-Nuernberg
- Prof. Cathleen Kantner, Stuttgart University
- Prof. Gerik Scheuermann, Leipzig University
- Prof. Maciej Piasecki, Wroclaw University of Science and Technology
- Prof. Chris Biemann, University of Hamburg
- Prof. Hiroshi Yoshida, Ritsumeikan University
- Prof. Verena Klemm, Leipzig University
- Dr. Stefan Jänicke, Leipzig University
- Mr. Peter Chan, Stanford University
- Prof. Martin Potthast, Leipzig University
- Prof. Shigeo Sugimoto, Tsukuba University
- Prof. Geoffery Rockwell, University of Alberta
- Mr. Peter Mühleder, Leipzig University Library
- Mr. Christian Kahmann, Leipzig University