

# Towards an encoding standard for social media and CMC:

## Experiences from German and French corpus projects using TEI

Michael Beißwenger, Thierry Chanier, Eric Ehrhardt,  
Axel Herold, Harald Lungen, Céline Poudat,  
Angelika Storrer, Ciara R. Wigham

in connection with Adrien Barbaresi, Alexander Geyken,  
Marc Kupietz, Lothar Lemnitzer, Andreas Witt



International Research Days:  
Social Media and CMC Corpora for the eHumanities



Journées Internationales de recherche  
«Médias sociaux et corpus de communication médiée  
par les réseaux. Annotation, analyse, données libres»  
23-24 octobre 2015

# Towards a TEI encoding standard for CMC

## 15 mins **Introduction** (Beißwenger, Chanier)

- Why do we need an annotation standard for CMC/social media?
- The “philosophy” of encoding textual genres with TEI (in a nutshell)
- Introduction of the TEI special interest group on CMC
- Overview: The challenge of modeling CMC in TEI ... and first suggestions for solutions (from CoMeRe and CLARIN-D)

## 15-20 mins **The CoMeRe French CMC corpora and their modeling in TEI** (Chanier, Poudat, Wigham)

Presentation and discussion of the TEI schema used for encoding the CMC corpora in CoMeRe

## 15-20 mins **Schemas and experiences from modeling German CMC corpora in TEI** (Beißwenger, Herold, Lungen, Storrer *et al.*)

Focus: TEI schema used for encoding the corpus data of the CLARIN-D curation project *ChatCorpus2CLARIN*

## 5 mins **How to use TEI schemas for corpus annotation** (Lungen)

## 15 mins **Discussion**

# Why do we need an annotation standard for CMC?

## Create your own, unique XML schema (eHumanities “1.0”)



schema perfectly fits with the needs of the individual project



schema is idiosyncratic, resource (corpus) is not interoperable with other resources



VS

## Comply with a standard (eHumanities “2.0”)



- facilitates the building of corpora (availability of schemas, best practices, and tools)
- sustainability of resources
- interoperability of resources (with corpora of the same type and with corpora of other types)

⇒ **Advanced opportunities for empirical research**



compliance with existing standard restricts the freedom to design everything in a way that perfectly fits for the peculiarities of CMC discourse

# What is and what does the TEI offer?

Annotation framework provided by the **Text Encoding Initiative (TEI)**: *De-facto* standard in the field of Digital Humanities:



[www.tei-c.org](http://www.tei-c.org)

- widely used interchange format for a variety of genres and document types (1st version of the TEI guidelines: 1990) ⇒ interoperability of resources
- A range of corpora/language resources are already represented in TEI.
- The TEI framework allows for a flexible adaptation and extension to new genres and document types which are not yet covered by the existing version of the standard.
- Very lively community organized in several special interest groups and workgroups which are continuously developing solutions for adapting the guidelines to new usage contexts and genres.



# What is and what offers the TEI?

Create your own, unique XML schema (eHumanities “1.0”)

VS

Comply with a standard (eHumanities “2.0”)



## customization

“Because the TEI Guidelines must cover such a broad domain and user community, it is essential that they be **customizable**: both to permit the creation of manageable subsets that serve particular purposes, and also to permit usage in areas that the TEI has not yet envisioned.”

## standardization

In view of the increasing importance of CMC as well as of the diverse needs to store and represent CMC data in corpora, a core framework for the representation of CMC genres should become part of the standard (which then can be customized for the specific needs of specific projects).

# The philosophy of encoding textual genres in TEI: The ‘pizza model’



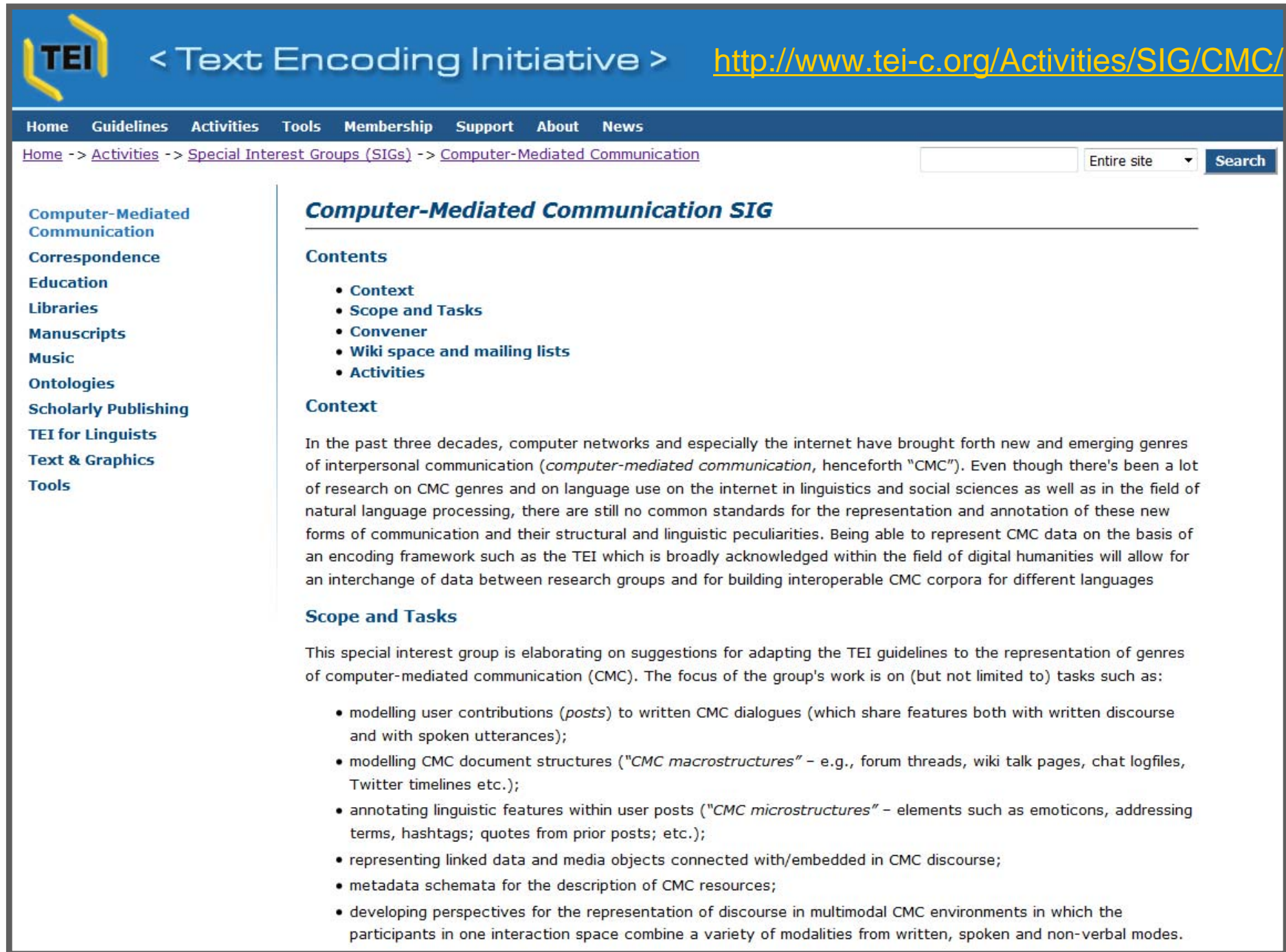
To build a TEI pizza, take...

- **the *pizza dough and base*** (tomatoes, cheese)  
= the basic structure of any TEI schema defined in the four modules *core*, *tei*, *header*, *text structure*
- ***toppings for specific pizza types*** (napoli, diavolo)  
= additional modules for particular text genres – e.g.:
  - transcribed speech
  - dictionaries
  - manuscripts
  - performance texts
  - dictionaries
- ***extra toppings of your choice*** (onions, anchovies)  
e.g., additional elements and attributes for specific concepts defined through customized modifications and extensions of the standard models

+ **computer-mediated communication ?**

(...)

# TEI Special Interest Group (SIG) on CMC (since 2013)



The screenshot shows the TEI website's navigation and content for the Computer-Mediated Communication SIG. The header includes the TEI logo and the text '< Text Encoding Initiative >' with a URL: <http://www.tei-c.org/Activities/SIG/CMC/>. A navigation menu lists: Home, Guidelines, Activities, Tools, Membership, Support, About, News. Below the menu is a breadcrumb trail: Home -> Activities -> Special Interest Groups (SIGs) -> Computer-Mediated Communication. A search box is present with a dropdown menu set to 'Entire site' and a 'Search' button. The left sidebar contains a list of categories: Computer-Mediated Communication, Correspondence, Education, Libraries, Manuscripts, Music, Ontologies, Scholarly Publishing, TEI for Linguists, Text & Graphics, and Tools. The main content area features the title 'Computer-Mediated Communication SIG' and a 'Contents' list: Context, Scope and Tasks, Convener, Wiki space and mailing lists, and Activities. The 'Context' section contains a paragraph about the history of CMC and the need for TEI standards. The 'Scope and Tasks' section lists specific goals of the SIG, such as modelling user contributions, document structures, linguistic features, linked data, metadata, and perspectives on multimodal CMC.

**TEI** < Text Encoding Initiative > <http://www.tei-c.org/Activities/SIG/CMC/>

Home Guidelines Activities Tools Membership Support About News

Home -> Activities -> Special Interest Groups (SIGs) -> Computer-Mediated Communication

Computer-Mediated Communication  
Correspondence  
Education  
Libraries  
Manuscripts  
Music  
Ontologies  
Scholarly Publishing  
TEI for Linguists  
Text & Graphics  
Tools

## Computer-Mediated Communication SIG

### Contents

- Context
- Scope and Tasks
- Convener
- Wiki space and mailing lists
- Activities

### Context

In the past three decades, computer networks and especially the internet have brought forth new and emerging genres of interpersonal communication (*computer-mediated communication*, henceforth "CMC"). Even though there's been a lot of research on CMC genres and on language use on the internet in linguistics and social sciences as well as in the field of natural language processing, there are still no common standards for the representation and annotation of these new forms of communication and their structural and linguistic peculiarities. Being able to represent CMC data on the basis of an encoding framework such as the TEI which is broadly acknowledged within the field of digital humanities will allow for an interchange of data between research groups and for building interoperable CMC corpora for different languages

### Scope and Tasks

This special interest group is elaborating on suggestions for adapting the TEI guidelines to the representation of genres of computer-mediated communication (CMC). The focus of the group's work is on (but not limited to) tasks such as:

- modelling user contributions (*posts*) to written CMC dialogues (which share features both with written discourse and with spoken utterances);
- modelling CMC document structures ("*CMC macrostructures*" – e.g., forum threads, wiki talk pages, chat logfiles, Twitter timelines etc.);
- annotating linguistic features within user posts ("*CMC microstructures*" – elements such as emoticons, addressing terms, hashtags; quotes from prior posts; etc.);
- representing linked data and media objects connected with/embedded in CMC discourse;
- metadata schemata for the description of CMC resources;
- developing perspectives for the representation of discourse in multimodal CMC environments in which the participants in one interaction space combine a variety of modalities from written, spoken and non-verbal modes.

# TEI Special Interest Group (SIG) on CMC

## Activities / “road map”:

- Feb. 2013: International workshop on CMC Corpora (Dortmund): Formation of the SIG
- June 2013: Work meeting at U Clermont/FR
- Sept. 2013: Special topic panel and 1st SIG meeting as part of the TEI Conference in Rome
- Feb. 2014: 2nd SIG meeting as part of the Empirikom conference on social media corpora, Dortmund
- Sept. 2014: 3rd SIG meeting as part of the DARIAH VCC Meeting, Rome
- Oct. 2015: Special topic panel and 4th SIG meeting as part of the TEI Conference in Lyon

**2016:** **Submit a proposal for CMC mdels to the TEI**



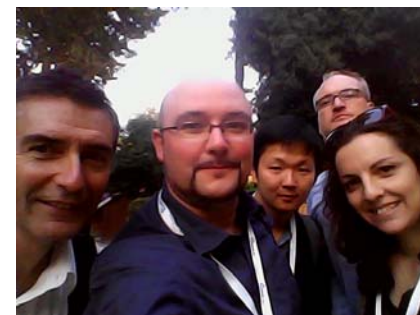
Empirikom  
Conference on Social  
Media Corpora,  
Dortmund, Feb. 2014



TEI Conference, Rome, Sept. 2013



DARIAH meeting, Rome, Sept. 2014





# TEI Special Interest Group (SIG) on CMC

## Working mode / schema drafts:

- 2012      1st draft for a TEI schema for CMC developed as part of the preliminary work for the DeRiK project («**DeRiK schema**»: Beißwenger, Ermakova, Geyken, Lemnitzer, Storrer 2012).
- 2013/14    Testing of schema draft 1 for use with the Wikipedia corpus in DeReKo (Margaretha & Lungen 2014).
- Discussion of schema draft 1 at workshops & conferences.
- 2nd draft for a TEI schema: special focus on multimodal CMC («**CoMeRe schema**»: Chanier, Poudat, Sagot, Antoniadis, Wigham, Hriba, Longhi, Seddah 2014).
- 2015      3rd draft for a TEI schema (building on the models and experiences from DeRiK and CoMeRe): «**CLARIN-D schema**» (Beißwenger, Ehrhardt, Herold, Lungen, Storrer).
- Discussion of models suggested in schema drafts 2+3 with the TEI community (⇒ **TEI-Conf. Lyon**) as well as with colleagues who are building CMC corpora (⇒ **IRD Rennes**).

# Schema drafts of the TEI-SIG on CMC

page discussion view source history watch

## Main Page

This is a wiki devoted to the [Text Encoding Initiative \(TEI\)](#). It is created by TEI-ers for TEI-ers, and if you wish to contribute something or join the discussions, you are most welcome – all you need to do is [login](#) or [register](#). Choose from the following:

### Documentation of schema drafts from the SIG

[http://wiki.tei-c.org/index.php/SIG:Computer-Mediated\\_Communication](http://wiki.tei-c.org/index.php/SIG:Computer-Mediated_Communication)

#### "CLARIN-D schema" (2015): TEI Schema and ODD from the CLARIN-D curation project *ChatCorpus2CLARIN*

**Project context:** The schema has been developed and tested with data from several CMC genres (chats, tweets, whatsapp, wikipedia talk pages, ...) as part of the work of the German CLARIN-D curation project *ChatCorpus2CLARIN*.

**Authors:** Michael Beißwenger, Eric Ehrhardt, Axel Herold, Harald Lungen, Angelika Storrer.

**Main characteristics compared to previous schema drafts (CoMeRe, DeRiK):**

- Reduction of new elements through re-modeling some CMC-specific concepts from the previous schemas with „standard“ TEI (guiding principle: "reduce to the max": introduction of new models and modification of existing models only for concepts which are needed *in any case*; for everything else: definition of best practices for the use of existing models in TEI-P5)
- Definition of an interface to part-of-speech annotations (using `<w>` and `<phr>`)

**ODD / documentation of the schema:** see detail page: [SIG:CMC/CLARIN-D schema draft for representing CMC in TEI \(2015\)](#)

**Presentation / discussion of the CLARIN-D schema:** The schema will be discussed in two panels at the following conferences.

- TEI across corpora, languages and genres: Towards a standard for the representation of social media and computer-mediated communication.* Panel at the Annual Conference and Members Meeting of the Text Encoding Initiative 2015: "Connect, Animate, Innovate", Université Lumière, Lyon 2 (F), 29 October 2015 (organized by Michael Beißwenger & Thierry Chanier).
- Towards an encoding standard for social media and CMC: Experiences from German and French corpus projects using TEI.* Panel at the International Research Days: Social Media and CMC Corpora for the eHumanities, Université Rennes 2, Rennes (F), 23-24 October 2015 (organized by Michael Beißwenger & Thierry Chanier).

#### "CoMeRe schema" (2014): TEI schema and ODD from the CoMeRe network

**Project context:** The schema has been developed in the context of the French network [CoMeRe \(Communication médiée par les réseaux\)](#) and used for annotation of [several corpora of French CMC](#) (SMS, tweets, chat, weblogs, multimodal CMC, ...).

**Authors:** Thierry Chanier, Céline Poudat, Benoit Sagot, Georges Antoniadis, Ciara R. Wigham, Linda Hriba, Julien Longhi, Djamé Seddah.

**Main characteristics compared to the previous schema draft (DeRiK):**

- Introduction of an element `<prod>` for the representation of non-verbal acts
- (re-)definition of `<post>`, `<prod>` and `<u>` as models which may be combined within one interaction (= installation of one main result of the SIG meeting 2013 in Rome). => make the schema fit for multimodal CMC
- includes a metadata schema for CMC

**ODD / documentation of the schema:** see detail pages:

- [SIG:CMC/CoMeRe schema draft for representing CMC in TEI \(2014\)](#)
- [CMC/CoMeRe metadata schema draft for CMC](#)

**Article in the JLCL special issue on CMC corpora:**

# Modeling CMC in TEI: The challenge

**Fundamental challenge (1):** Written CMC shares characteristics both with *text* and *spoken conversation* ...

- o Just like *spoken conversation* (and different from *text*), CMC is dialogic interaction in which each communicative move creates/changes the context for follow-up moves.
  - o Just like *text documents* and different from spoken conversation, written CMC is organized through the exchange of stretches of written text which have completely been composed before they are transmitted and read.
- ⇒ A basic model for the representation of user contributions to written CMC („posts“) should reflect these properties.

**Fundamental challenge (2):** A basic schema for CMC should be flexible enough to represent also multimodal CMC interactions

- o It should include **models for the representation of non-verbal acts** – acts performed by the human body (mediated through webcams), by the simulated body of an avatar, acts performed through actions in groupware / shared editor tools (etc.)
- o It should allow for a representation of **interactions in which the participants combine written and spoken conversation with non-verbal acts** (e.g., in „virtual“ 3D worlds, in multimodal learning environments)

# TEI schema drafts of the SIG: Some basic features

## DeRiK schema (Beißwenger, Ermakova, Geyken, Lemnitzer, Storrer 2012):

- Introduction of an element model `<post>` for written user contributions to CMC interactions which combines features of text divisions and spoken utterances.
- Adaptation of the existing element model `<div>` for the representation of CMC threads and logfiles.
- Introduction of diverse models for CMC phenomena below the `<post>` level.

## CoMeRe schema (Chanier, Poudat, Sagot, Antoniadis, Wigham et al. 2014):

- Introduction of an element `<prod>` for the representation of non-verbal acts
- (re-)efinition of `<post>`, `<prod>` and `<u>` as models which may be combined within one interaction. ⇒ [make the schema fit for multimodal CMC](#)

## CLARIN-D schema (Beißwenger, Ehrhardt, Herold, Lungen, Storrer 2015):

- [Reduction of new elements](#) through re-modeling some CMC-specific concepts from the previous schemas with „standard“ TEI
  - ⇒ [guiding principle: „reduce to the max“](#): introduction of new models and modification of existing models only for concepts which are needed *in any case*; for everything else: best practices for the use of existing models
- Definition of an [interface to part-of-speech annotations](#)