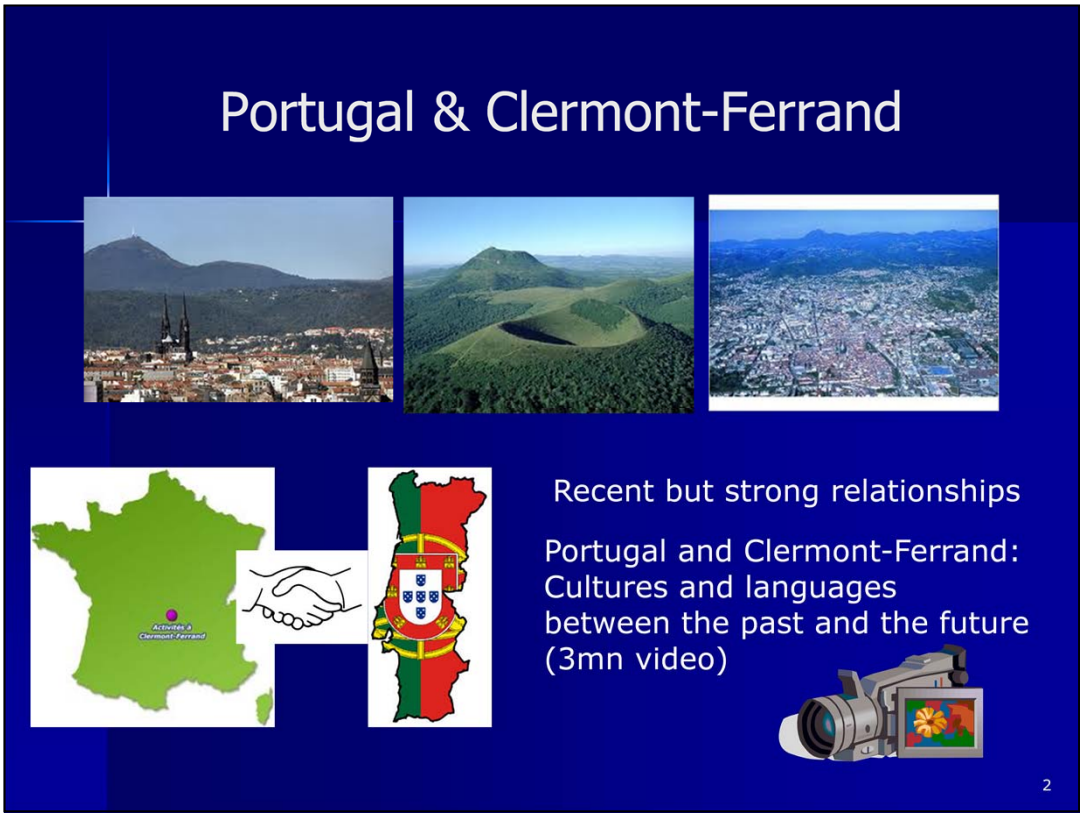
A viewpoint on the place of CALL within the Digital Humanities: considering CALL journals, research data and the sharing of research results

Thierry Chanier

Download slides and videos, extended. version online

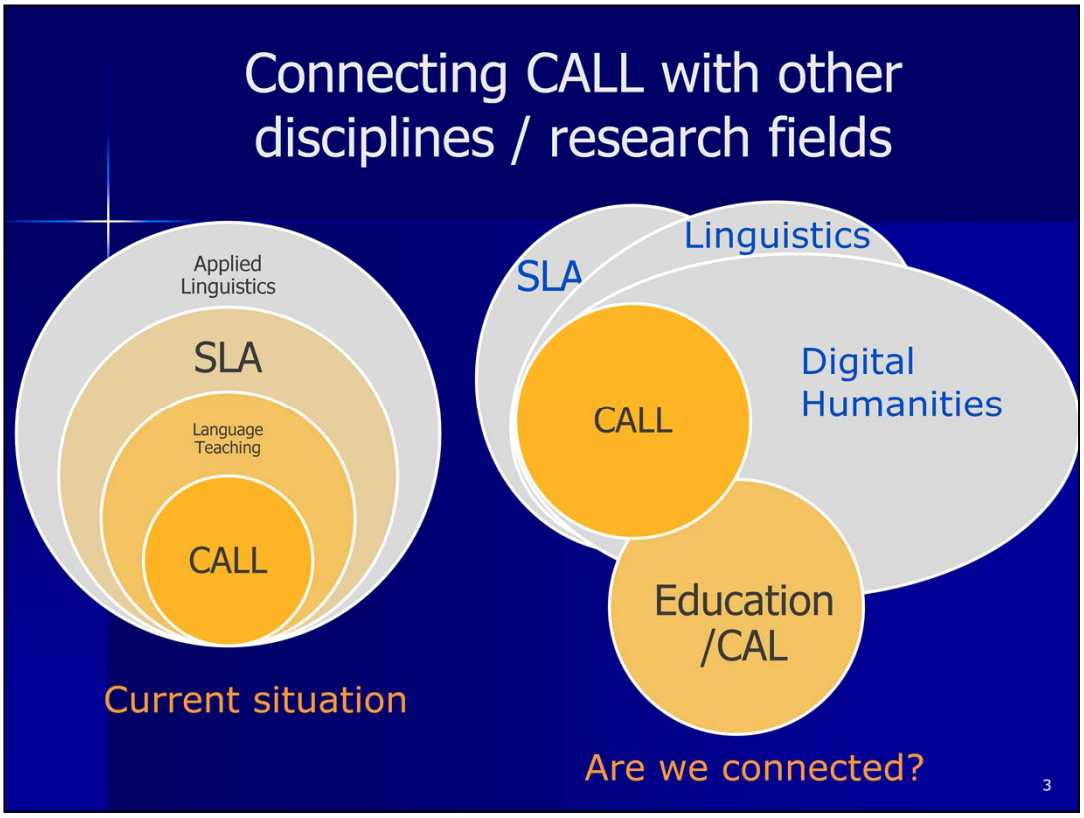


Caros colegas e amigos,

Não falo Português e nunca aprendi, más antes de eu começar a minha apresentação sobre o tema referente as revistas e dados sobre as pesquisas, eu gostaria de fazer uma breve introdução sobre Portugal e minha cidade Clermont-Ferrand.

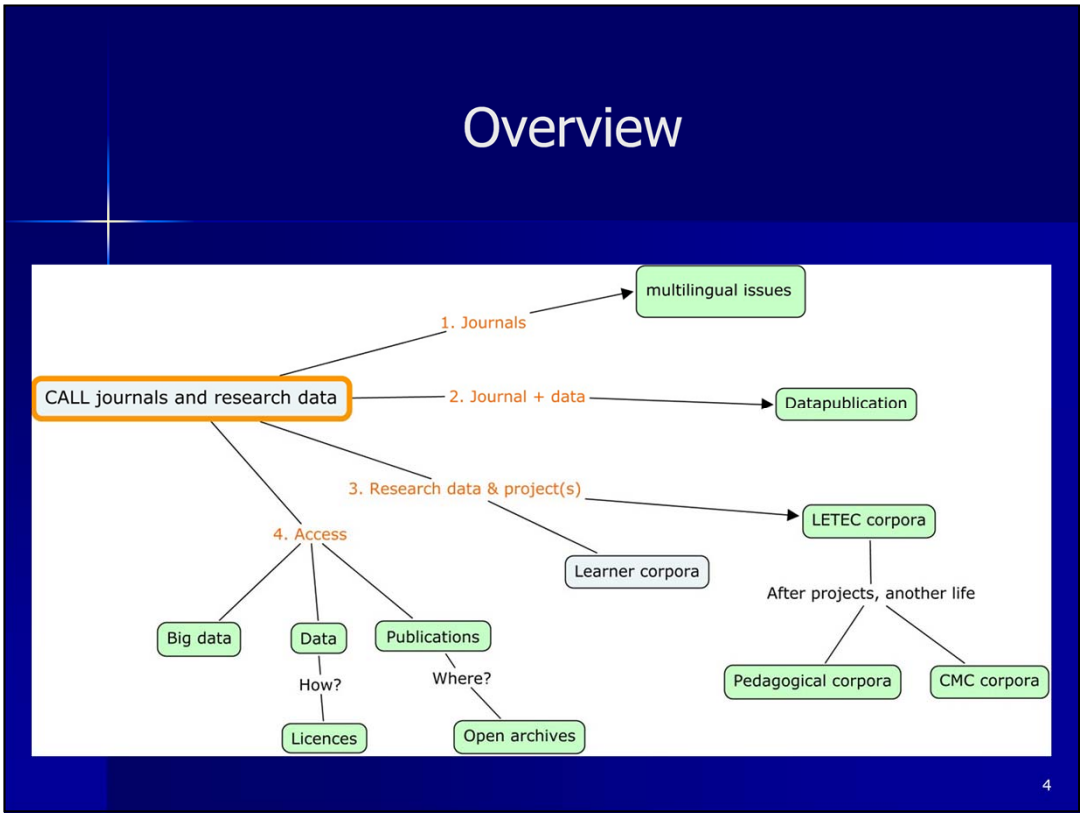
Esta cidade é a segunda maior cidade Portuguesa aqui na França, depois de Paris. Este pequeno vídeo mostra bem as relações que existem entre os nossos dois idiomas e culturas.

Obrigado pela vossa atenção.



Generally now when you read CALL literature or attend CALL conferences, our domain is presented as being a subpart of fields such as Language teaching, Second Language Acquisition, not to say Applied Linguistics.

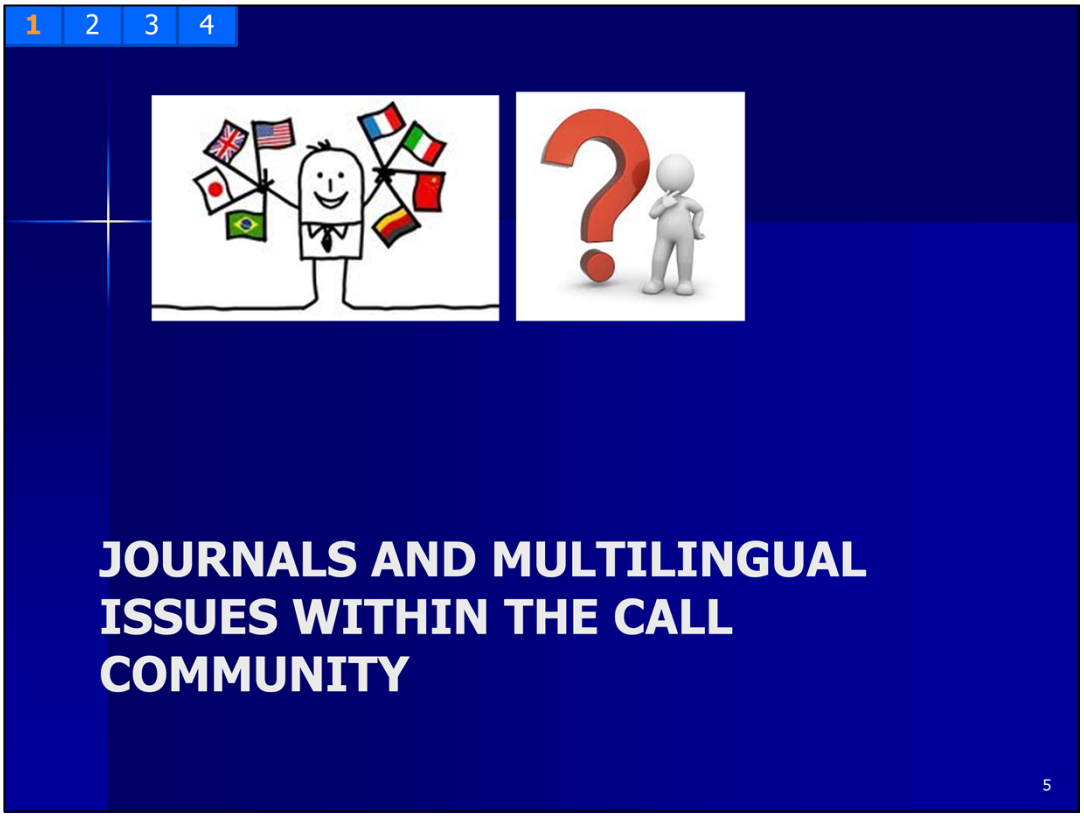
Without rejecting this standpoint, I would like here to turn our attention to other fields, more particularly to the so-called “Digital Humanities”.



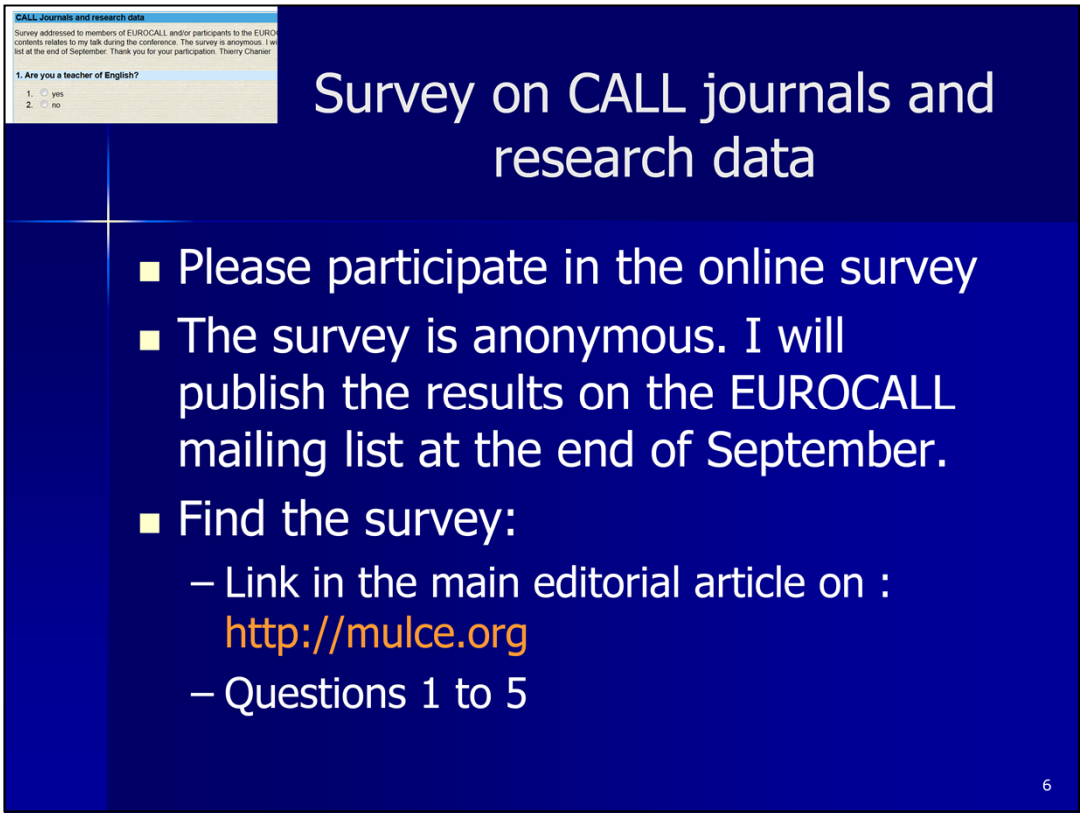
The term "Digital Humanities" (DH) made a buzz at the MLA (Modern Language Association) convention in 2009. The term is now in widespread use within the Humanities. CALL may be directly concerned: our field belongs to the Humanities and, from the outset, we have had a strong interest in computers and computing.

Although various meanings and interpretations can be attributed to this term, this presentation will address issues related to ways of promoting CALL research in order to meet what may soon become research standards within the Humanities.

We will talk about journals, data linked to publications, research data organized as corpora and of access to these publications and research data.



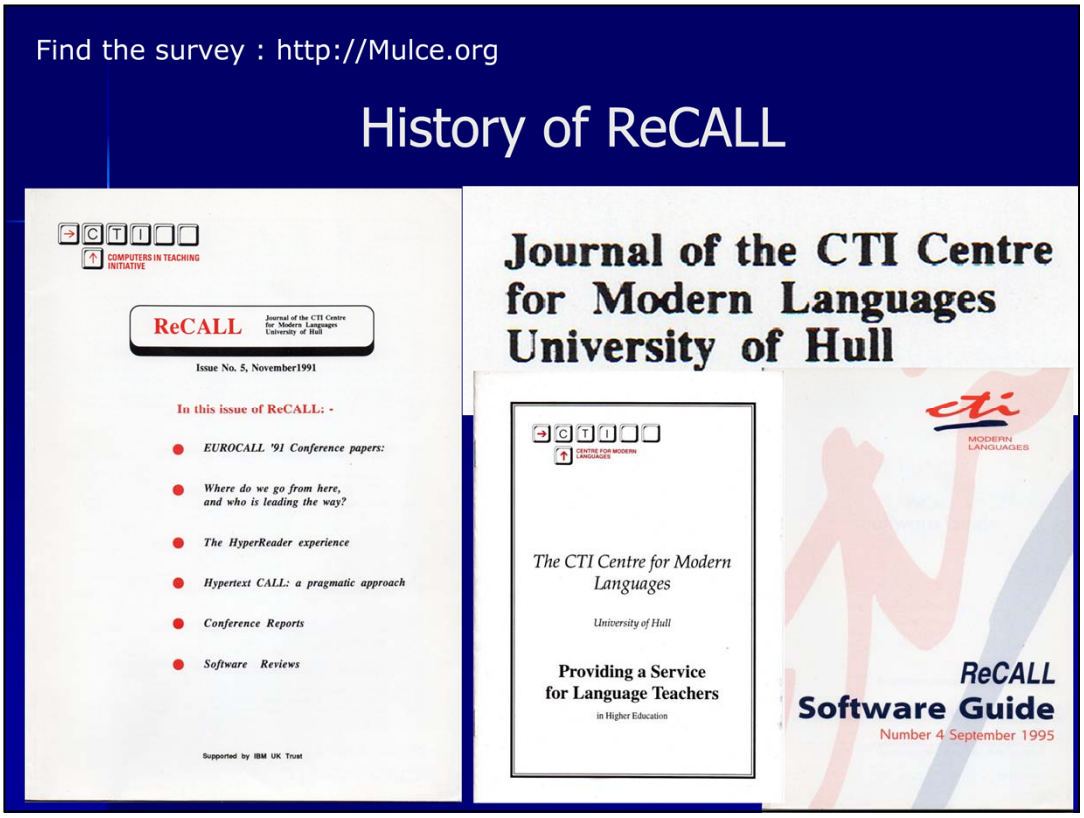
First of all, let us start with CALL journals (and conferences) and (briefly) examine whether we have built a multilingual academic community



About this first issue, as well as the other ones, I would be happy to collect, in an anonymous way (no possibility for me to know who answered what) you current standpoints.

If not already done before the conference you can access the survey. The link to the survey is available on the front page of the site mulce.org.

Of course I will publish within one or 2 weeks results on Eurocall and CALICO memberlists.



Since it is the 20th anniversary of Eurocall, let me first remind you that our journal ReCALL has got a long tradition. It first appeared in 89, and was at that time supported by the CTI centre for Modern Languages. At the end of the eighties, was launched a British national initiative to support the development of technology, CAL in education. The university of HULL, where Graham Chesters and June Thompson were working, was in charge of CALL development.

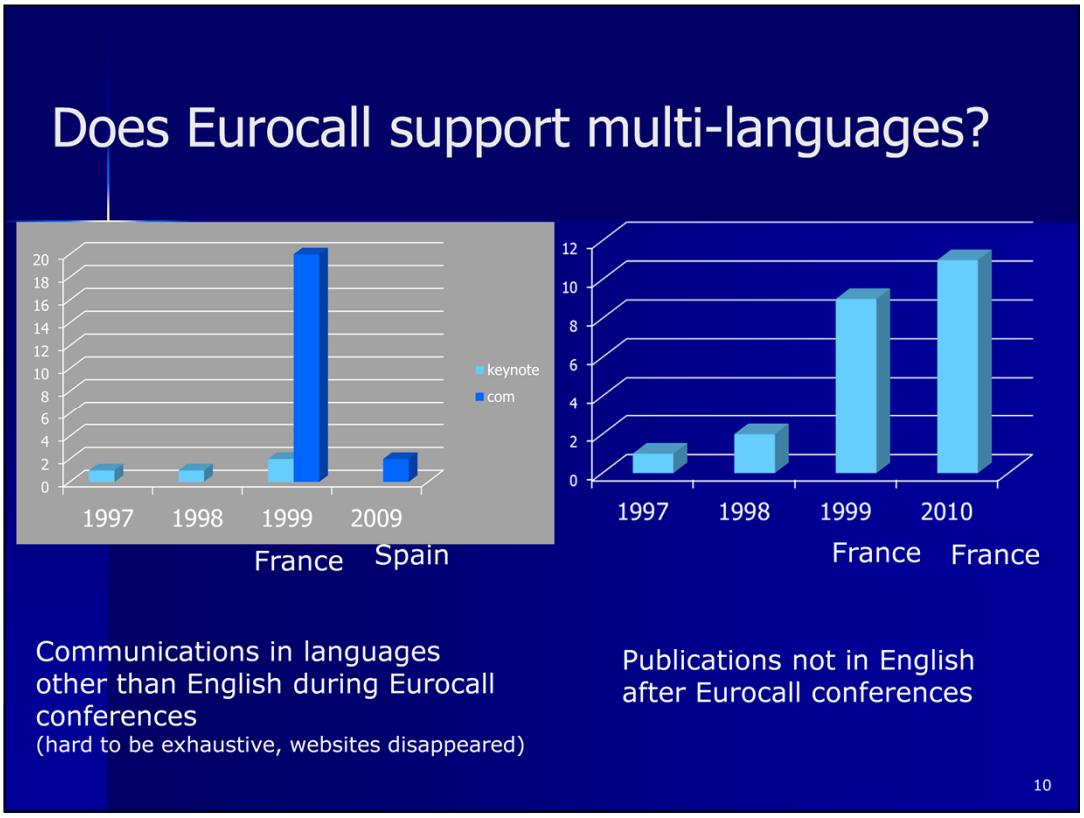
Besides publishing the Recall journal, Hull also provided a lot of services to language teachers. Teachers could visit the CTI and get software demos, or people from the CTI could come to their own institutions. The CTI also regularly published a software guide, and abroad we always were waiting for the last release.

In 95, ReCALL became a joint publication of the CTI and the association Eurocall. In 2003, Eurocall and Hull who still had the editing responsibility gave the publishing task to Cambridge University Press. Of course I would not like to omit to cite the Eurocall review, edited par Ana Gimeno, review still directly published by Eurocall.

Since the beginning of Recall there has been one person who still is in charge of ReCALL, namely June Thompson. Unfortunately for the first time , June could not attend our conference, but we could here send her a real thanks for all the wonderful work she achieved !

In 95, ReCALL became a joint publication of the CTI and the association Eurocall. In 2003, Eurocall and Hull who still had the editing responsibility gave the publishing task to Cambridge University Press. Of course I would not like to omit to cite the Eurocall review, edited par Ana Gimeno, review still directly published by Eurocall.

Since the beginning of Recall there has been one person who still is in charge of ReCALL, namely June Thompson. Unfortunately for the first time , June could not attend our conference, but we could here send her a real thanks for all the wonderful work she achieved !



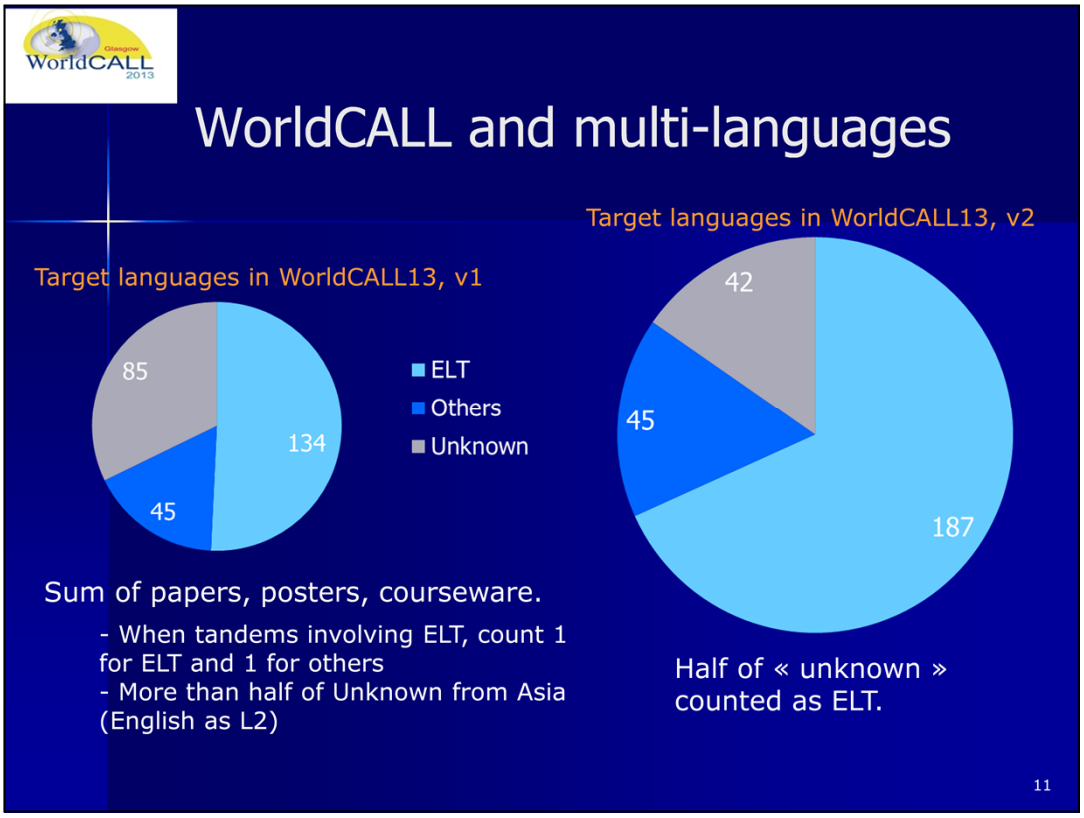
Now let us consider the multilingual issue. I made a quick enquiry among the discontinuous information available on Eurocall websites concerning communications not given in English during our conferences (starting in 93 up to 2012).

I could only note one keynote given in 97 and one in 98.

In 99 in Besançon we had one keynote and nearly 20 communications not in English, 2 in 2009 in Spain.

When looking at papers, full papers, published after Eurocall conferences, I found one in 97 in Recall, 2 in 98, but 8 and 11, respectively after the 99 and 2010 conferences in France.

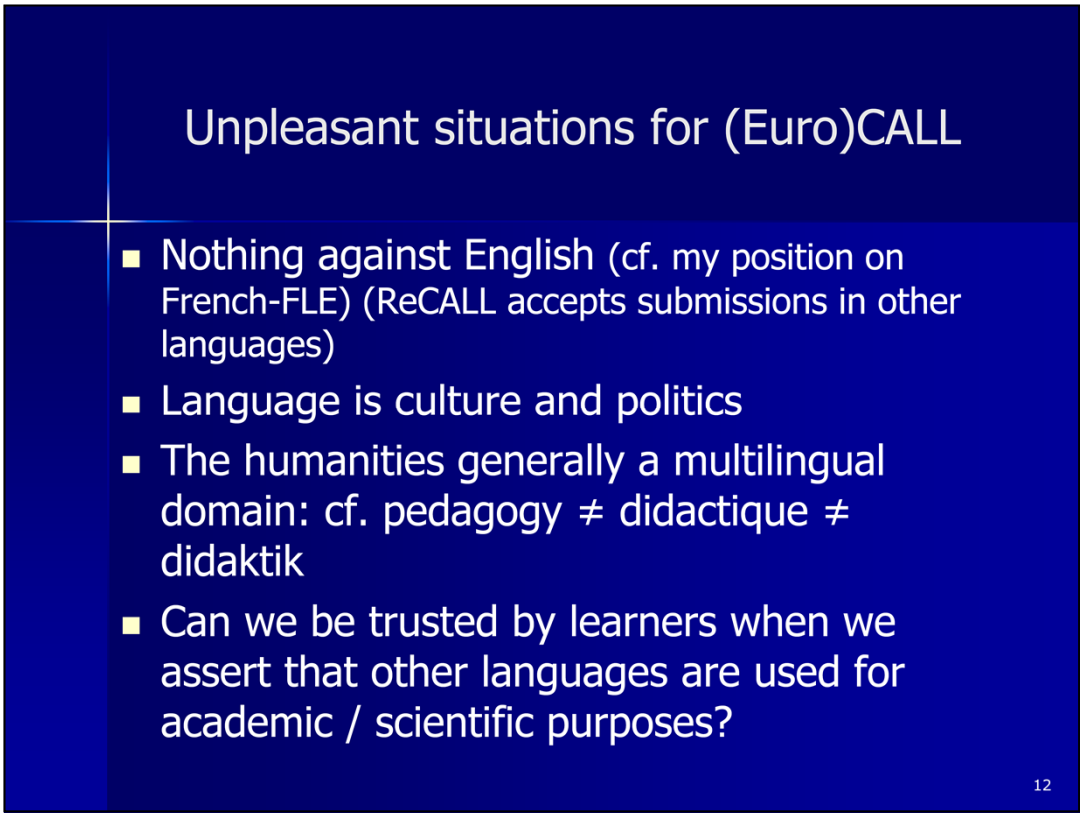
As you can see, the whole amount of communications and articles not in English is fairly reduced.



Let us have a look outside Eurocall. In July this year we had the WorldCALL conference in Glasgow. There were more than 250 events, whether papers, posters, courseware. 134 of them were given by English language teachers and, naturally, only related to this language. Only 45 events were produced by colleagues teaching other languages. 85 authors did not mention any language in their abstracts.

When looking more closely to this latter 85 events, a majority of them were organized by colleagues from East Asia. A very large majority of them are teachers of English. So I can reasonably count half of this 85 events as being connected to ELT. This gives us around 190 ELT events. In other words 70% of the conference related to ELT, versus only 17% clearly not referring to the English language.

Without denying a good scientific quality to WorldCALL conferences, one may wonder whether we are not building another kind of TESOL conference.



Please do not misunderstood my viewpoint. I have nothing against English. When we created the first French speaking CALL journal, namely Alsic, we did it in full collaboration with the other English speaking journals. We did jointly organized Eucall99 in Besançon.

Neither can we complain against ReCALL. We always have welcomed submissions of papers in other languages.

But we all know that language issues are closely connected to cultural and political ones. Hence we may be on the way of losing our autonomy.

Moreover, Humanities is generally considered by other scientific fields as a place where researchers can publish good quality papers in vernacular languages. What would it mean for example to be obliged to translate the words “didactique” in French, or “didaktik” in German into pedagogy when knowing these 3 terms have each a distinctive meaning representative of different perspectives in the Educational Research field?

Eventually, can we really be trusted by our learners when we try to explain them “Oh yes please learn Arab, or French, or Spanish, Portuguese, etc. because these are important scientific languages” !

What can we do in order to develop multilingual spaces in our community:

Develop international CALL journals

After Eurocall99 (Besançon) publications in ReCALL and in Alsic

After Eurocall2010 (Bordeaux)publications in ReCALL and in another journal

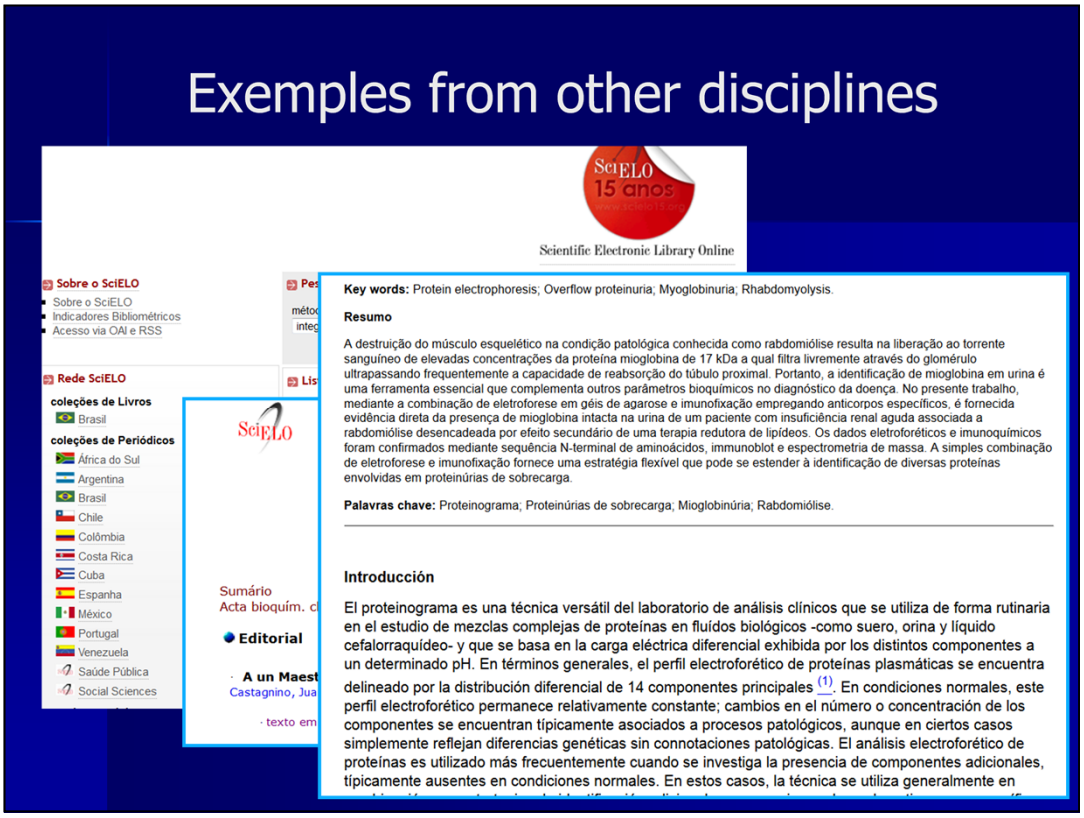
14

I mention the larger number of papers published in other languages than English after these two Eurocall conferences organized in France. How did this happen?

In Besançon we offered the possibility to authors to submit papers either to ReCALL or to Alsic . In Bordeaux, proceedings have been published with another publisher.

These opportunities have been jointly organized with EuroCALL. ReCALL lost nothing.

How can we develop publications in other languages?



In fact, it already exists in other disciplines.

Here is the example of the very large scientific publisher named Scielo. It publishes scientific journals in Spanish and Portuguese with authors belonging to America and Europe. Here is an illustration of a journal in biochemistry published in Portuguese.

You could think “well this is only an example for natural science and technology”. Indeed in Humanities we have a great opportunity with the public publisher OpenEdition. It is managed by academic people, it already publishes 400 journals in the academic field. This is where, for exemple, the Alsic editorial board moved 5 years ago in order to publish its journal. It has now set up offices in Spane and Portugal, publishes in these languages (here a journal in Sociology) and start discussing with German editorial teams.

So, why not soon start publishing an international CALL journal in Spanish & Portuguese (only one for America & Europe) to ensure a strong scientific basis from the beginning?

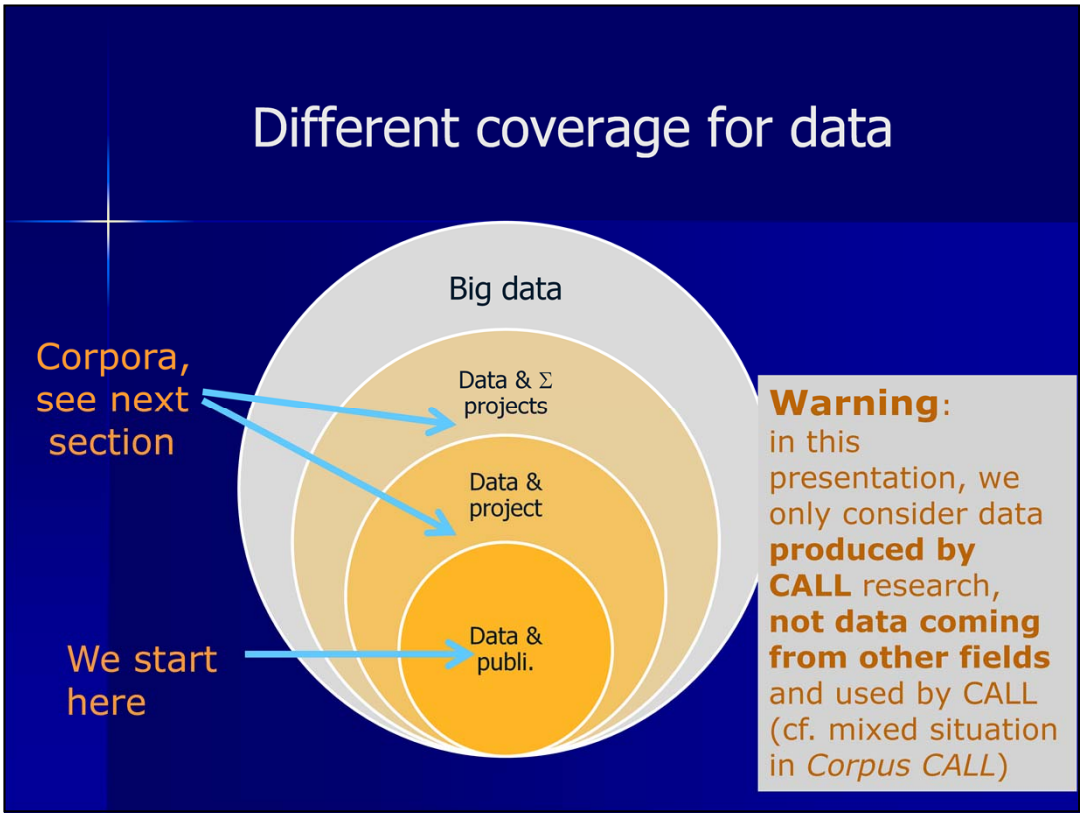
You could think “well this is only an example for natural science and technology”. Indeed in Humanities we have a great opportunity with the public publisher OpenEdition. It is managed by academic people, it already publishes 400 journals in the academic field. This is where, for exemple, the Alsic editorial board moved 5 years ago in order to publish its journal. It has now set up offices in Spane and Portugal, publishes in these languages (here a journal in Sociology) and start discussing with German editorial teams.

So, why not soon start publishing an international CALL journal in Spanish & Portuguese (only one for America & Europe) to ensure a strong scientific basis from the beginning?

You could think “well this is only an example for natural science and technology”. Indeed in Humanities we have a great opportunity with the public publisher OpenEdition. It is managed by academic people, it already publishes 400 journals in the academic field. This is where, for example, the Alsic editorial board moved 5 years ago in order to publish its journal. It has now set up offices in Spane and Portugal, publishes in these languages (here a journal in Sociology) and start discussing with German editorial teams.

So, why not soon start publishing an international CALL journal in Spanish & Portuguese (only one for America & Europe) to ensure a strong scientific basis from the beginning?

After having considered publications, let us turn our attention to research data



What types of data can we find in research?

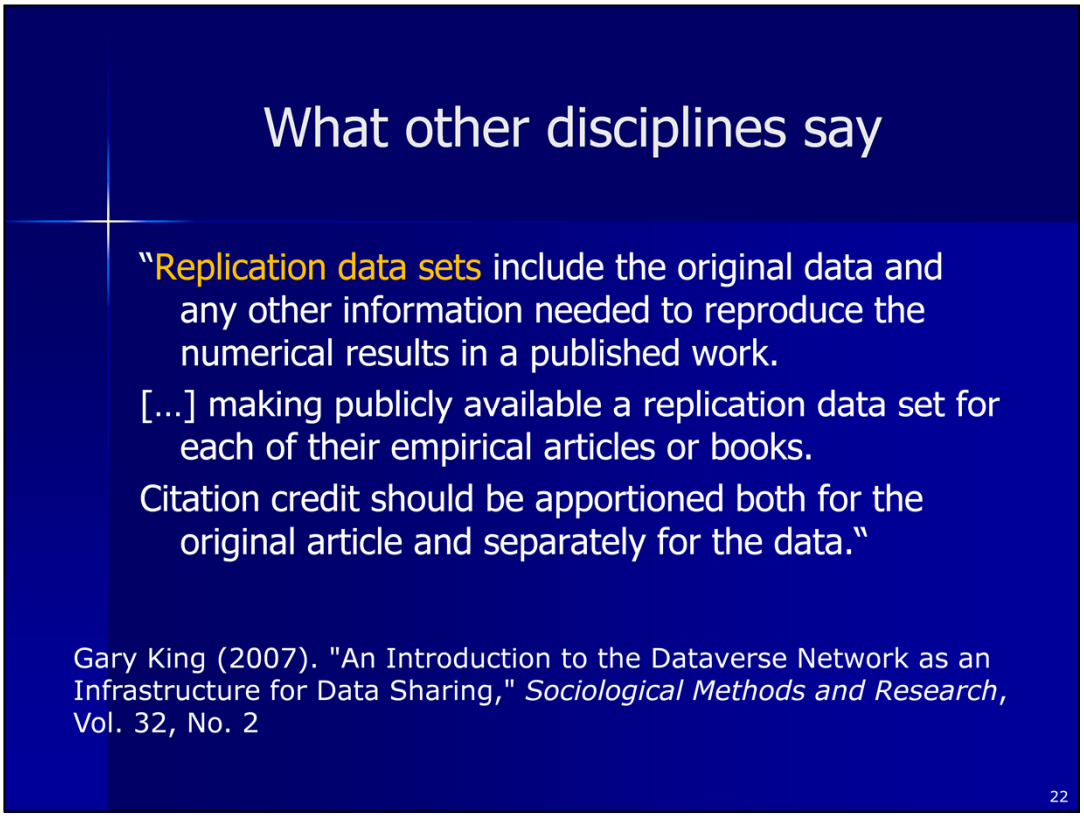
- Data linked to publications
- Data coming out of research projects (one or several CALL experiments)
- A recent new type the so-called Big Data (see the extented version)

Let us start with data linked to publications

Current situation in CALL (and many, but not all, fields in Humanities)

- Some (not all) of our papers are based on research data
- These data (empty forms, forms filled, spreadsheets, transcriptions, language data and their computation, audio, video, etc.) are not accessible to reviewers, nor to the readers once papers are published

21



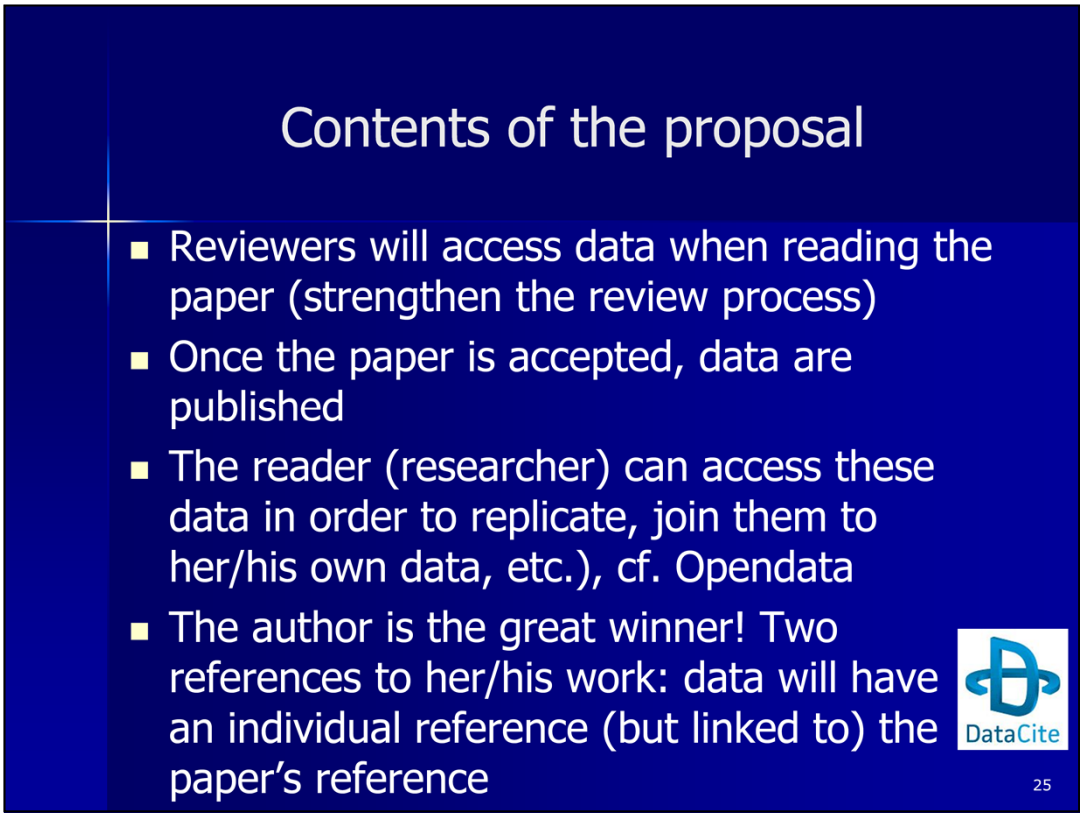
What other discplines in Humanities say about this?

Gary King, a sociologist, explains why we should make data, he called « replication data sets » available with publications.

What does Europe say?

A recent text published by the commission with respect to the new 2020 European framework proposes, in order to improve scientific communication, to develop a pilot on open access to data, primarily those data underlying scientific publications.

In CALL, since we are not waiting for the last train to arrive, we started discussing a joint project among 5 journals, European and North American ones, which have the habits of already cooperating, where part of the people here in the room are reviewers or members of the editorial boards.

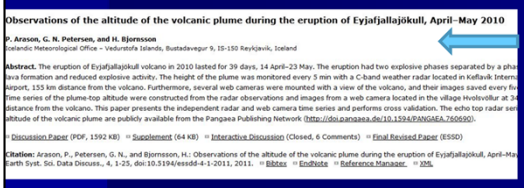


Here is the contents of the proposal made to these journals.

I am pleased to announce that yesterday the ReCALL editorial board, supported by CUP, accepted to joint the project !

Link between publication & data: example from earth sciences

Arason, P et al. (2011): Plume-top altitude time-series during 2010 volcanic eruption of Eyjafjallaj??l. Icelandic Meteorological Office, Reykjavik, doi:10.1594/PANGAEA.760690, Supplement to: Arason, Pordur; Petersen, G N; Bjornsson, H (2011): Observations of the altitude of the volcanic plume during the eruption of Eyjafjallajl, April-May 2010. *Earth System Science Data*, 3, 9-17, doi:10.5194/essd-3-9-2011


Observations of the altitude of the volcanic plume during the eruption of Eyjafjallajökull, April–May 2010

P. Arason, G. N. Petersen, and H. Bjornsson

Icelandic Meteorological Office – Veðurstofa Íslands, Bustadavegur 9, IS-150 Reykjavik, Iceland

Abstract. The eruption of Eyjafjallajökull volcano in 2010 lasted for 39 days, 14 April–23 May. The eruption had two explosive phases separated by a pha
lava formation and reduced explosive activity. The height of the plume was monitored every 5 min with a C-band weather radar located in Keflavík Intern
Airport, 155 km distance from the volcano. Furthermore, several web cameras were mounted with a view of the volcano, and their images saved every fiv
Time series of the plume-top altitude were constructed from the radar observations and images from a web camera located in the village Hvolsvöllur at 34
distance from the volcano. This paper presents the independent radar and web camera time series and performs cross validation. The echo top radar ser
altitude of the volcanic plume are publicly available from the Pangaea Publishing Network (http://doi.pangaea.de/10.1594/PANGAEA.760690).

Discussion Paper (PDF, 1592 KB) Supplement (64 KB) Interactive Discussion (Closed, 6 Comments) Final Revised Paper (ESSD)

Citation: Arason, P., Petersen, G. N., and Bjornsson, H.: Observations of the altitude of the volcanic plume during the eruption of Eyjafjallajökull, April–Ma
Earth Syst. Sci. Data Discuss., 4, 1-25, doi:10.5194/essdd-4-1-2011, 2011. BibTex EndNote Reference Manager XML

Journal site

Data site


PANGAEA

Data Publisher for Earth & Environmental Science

Data Description

Download Data

Datasets listed in this Collection

WHAT WE STARTED TO DO IN FRANCE

27

Datapublication (French project)

- With the help of TGE-Adonis (national infrastructure for humanities)
  - Now part of Huma-Num
- For Alsic and Sticef journals (as a starting point)
- Every journal has its entries, have an internal review process (cf. OJS) for data
- Reviewers can look at data when reading the paper (data are not open at this stage)
- When the paper is accepted data are published

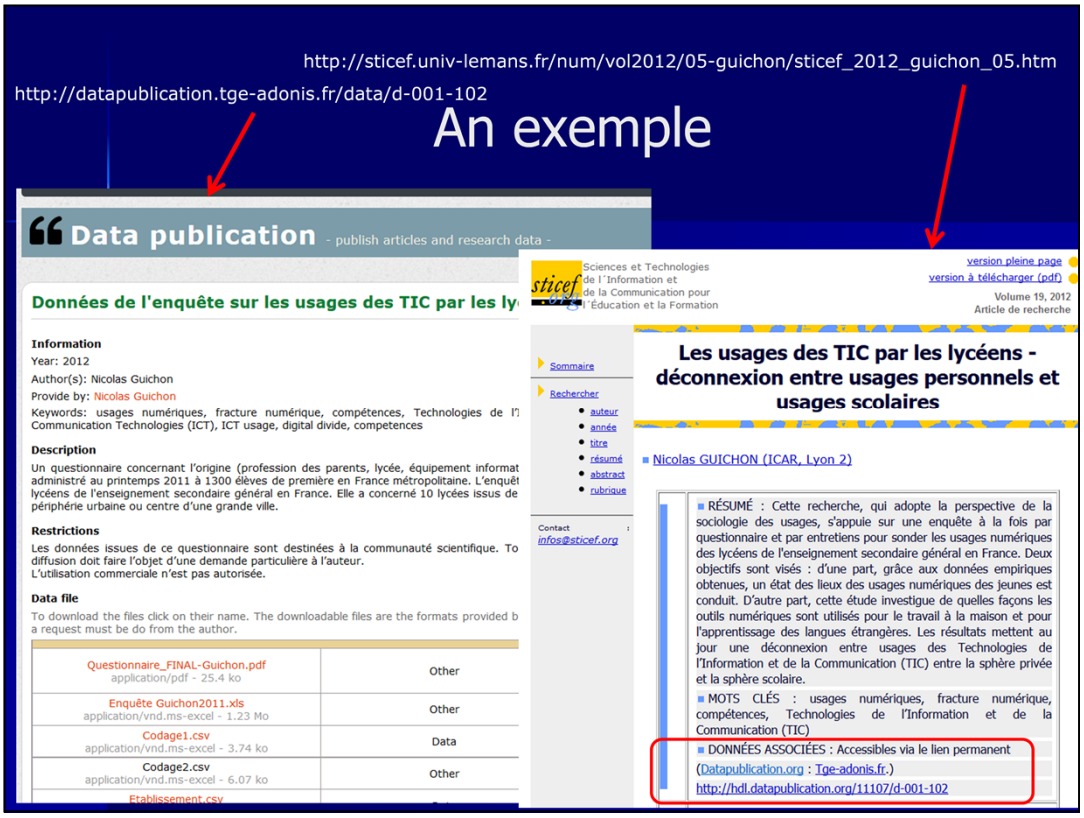
http://datapublication.tge-adonis.fr

28

http://sticef.univ-lemans.fr/num/vol2012/05-guichon/sticef_2012_guichon_05.htm

http://datapublication.tge-adonis.fr/data/d-001-102

An exemple



IRIS IS NOT THE PROJECT WE ARE LOOKING AT

30

http://www.iris-database.org

Why not IRIS?

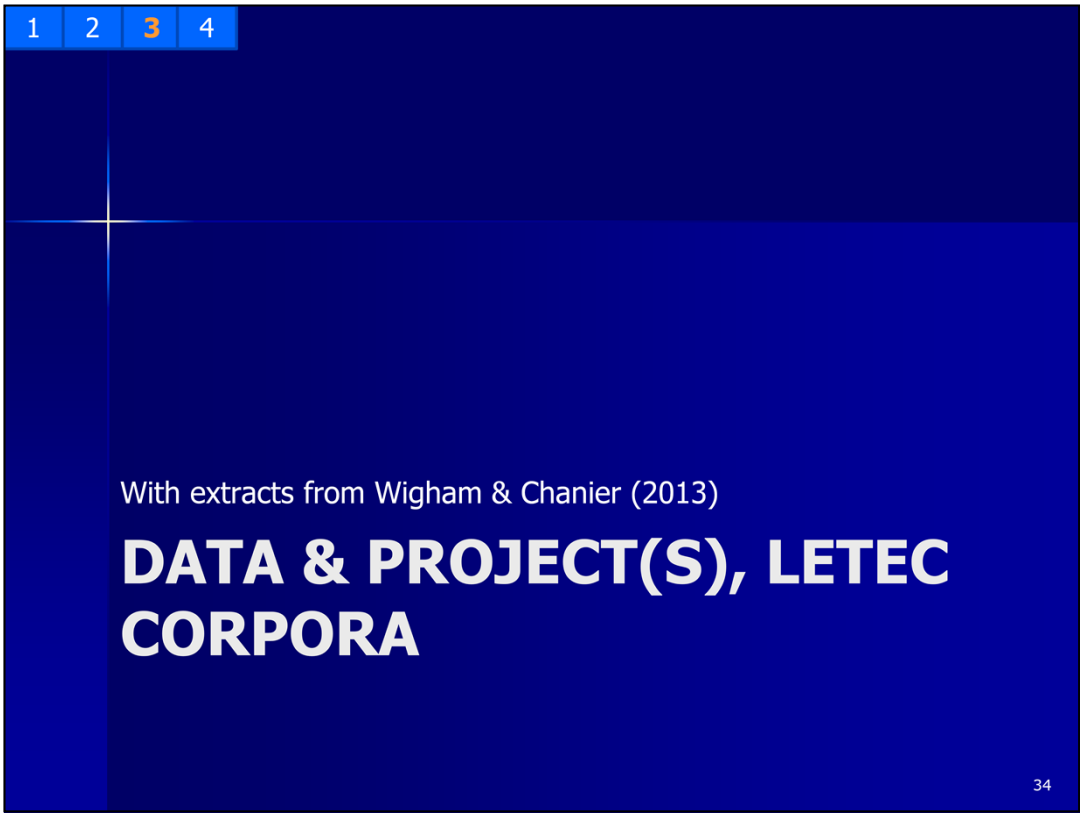
- Iris is an interesting OpenData project with links to journals from UK and USA universities, sponsored by UK, but …
- Data are not part of the review process
- Once a paper is accepted authors do as they pleased, e.g: some put the form of a questionnaire, not the data collected (answers), nor the computation (spreadsheet)
- Metadata are not standard (just for search on the site, like Merlot)
- They are local and cannot be harvested
- No reference to the data (cf. DataCite) , no permalink
- No crosslink between data and publication (which would not make sense because data are not exhaustive) and have not been part of the evaluation process

32

CALL Datapublication project

- Make a common proposal at the European Union level (Resear agency) via DARIAH
- Get logistical and official scientific support in order to design and open a website site (Datapublication)
- Where our 5 journals will have separate access for their editorial board in order to manage distinct review process
- Manage a joint design for the workflow of the review process
- Metadata format will be standard, permalink given, full reference with link and full reference of papers
- When the web site is open, for every journal author's guidelines need to be changed (when authors submit papers which rely on data) and links be implemented in order to point from the journal to the data site
- Then the Datapublication website may be open for other journals in humanities (best to get EU support) whether they are based in or out of EU

33



In this 3rd section we will now consider data stemming from research projects.

First corpora in CALL : learner corpora

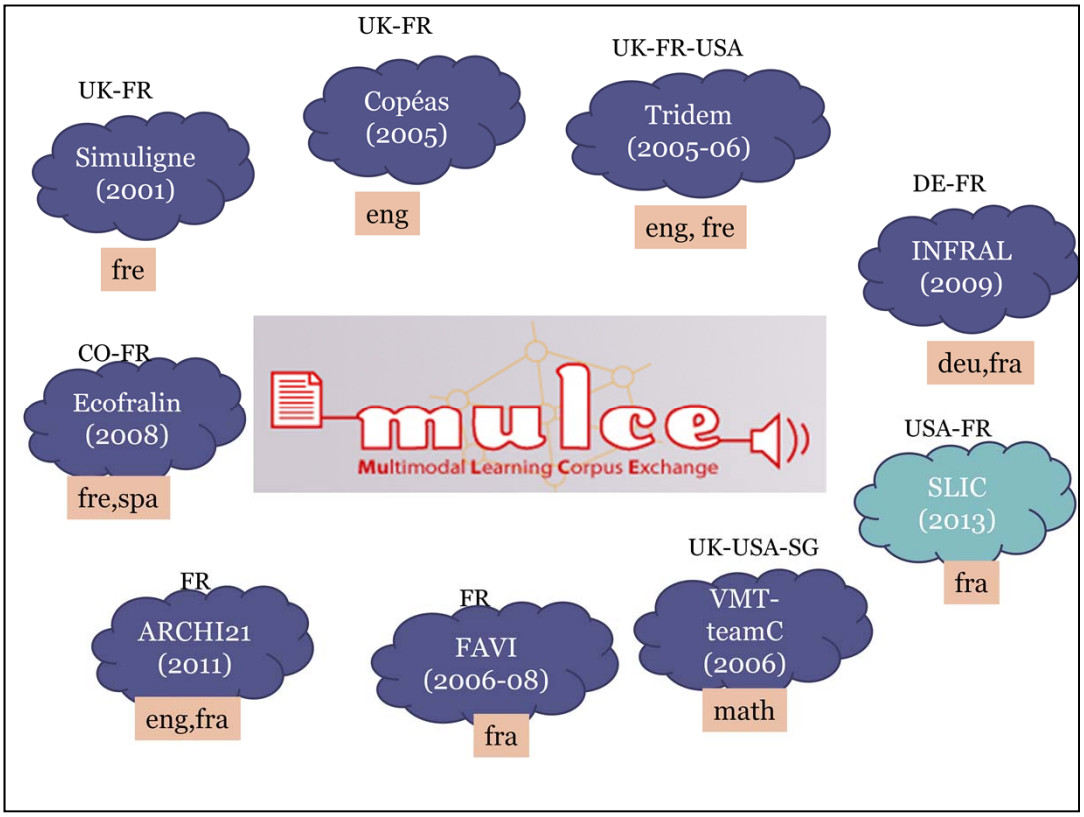
- Building corpora : collecting learners'production (essais), structuring, annotating, processing
- Using corpora
  - To enhance learning (DDL: data driven learning) under some circonstances
  - To enhance research
- Thinking about : Eurocall SIG, conferences, special issues, etc.

35

Generally when one talks in CALL about data assemble into corpora, we immediately think of learner corpora.

As you know, a learner corpus is made out of learners' productions. It is studied as a way to enhance language learning or the understanding of the learning process. Since there are plenty of opportunities on this issue in this conference, I will not say more on learner corpora, except when mentioning the question of access to data.

Here I would like to introduce another type of corpus, the Learning & teaching corpus (LETEC). Examples will be taken from the field of online learning situations in a multimodal context.



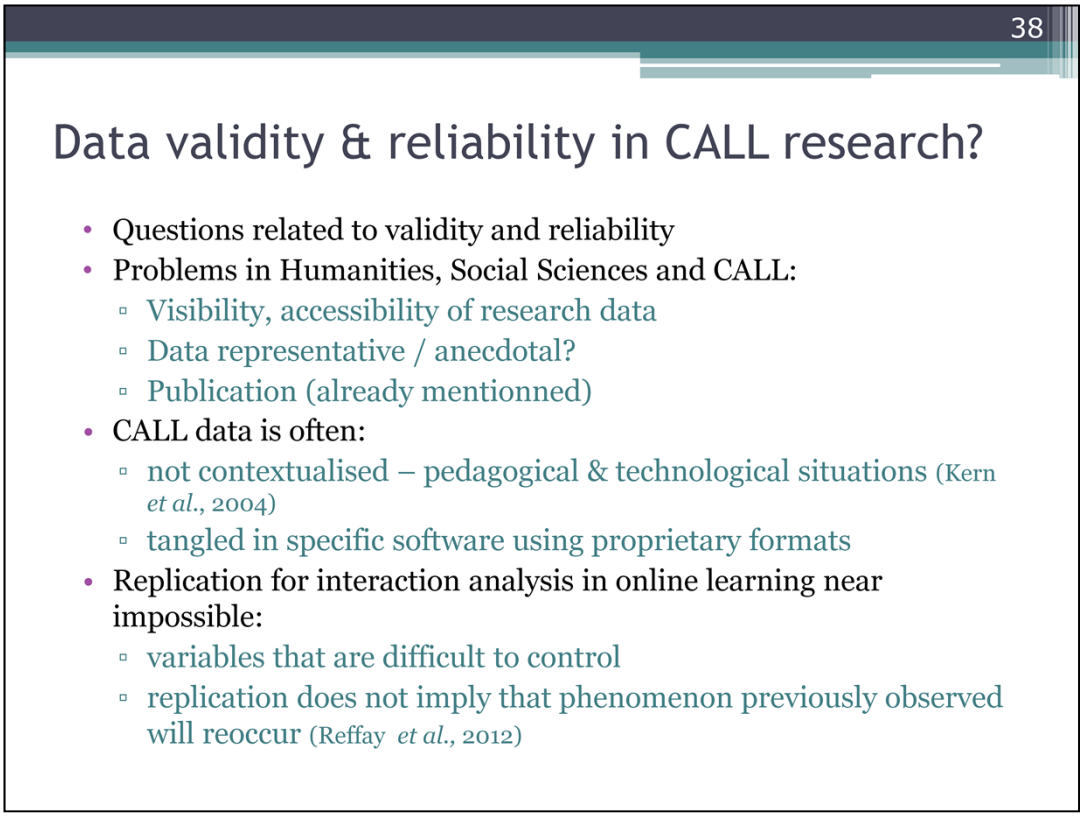
Over the past 12 years a community of researchers have been involved in online language learning projects, either for designing pedagogical scenarios, research protocols, for online tutoring, for collecting, analysing data or/and for publishing.

Projects started in 2001, with the global simulation Simuligne. From 2007 some of us decided that it was time to adopt a coherent and systematic way to organize data in order to improve our research methodology for reasons I will soon mention. It gave birth to the Mulce project.

It is impossible for me to cite all my colleagues here. Let me just mention Christophe Reffay who was there from the beginning, Marie-Noelle Lamy with whom we co-develop the Simuligne project and the Mulce project. Marie-Laure Betbeder, Maud Ciekanski and Ciara Wigham came later on and made a great deal of work. Chris Jones, present here,

co-constructed the Tridem project with Mirjam Hauck, Tim Lewis,, and Bonie Youngs.

As you can see here, every cloud corresponds to an online learning situation, a research project. Above clouds you have IDs of the country involved (Colombia, Germany, UK, USA, etc.), most of the situations correspond to what is now called a telecollaborative project. Under the clouds you have the languages that were at stake.



Mulce researchers were concerned with questions related to validity and reliability. When we set up a learning situation, study it and publish to which extent what we say is anecdotal or can be generalized? Did we actually studied what we pretended to study or did we neglected hidden factors which may open the way to other explanations, conclusions?
When we want to give some insurance about these issues and discuss them with other researchers, we are in a real trouble, for many reasons.
Among others:

- As mentioned no data associated with publications
- Almost no research data accessible nor visible
- When you do have slices of data they often are not contextualized (what were the precise technological and pedagogical situations?)
- They are tangled in specific software using proprietary formats

Research data quality: Mulce project

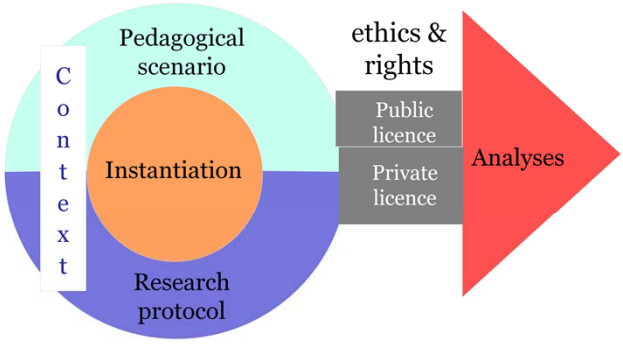
- Interoperability:
  - Structured and coherent data sets

  => analyses can be completed by researchers who did not participate in the course
- Sustainability:
  - Independent from online platforms
  - Stored in independent formalisms
- Open access to research data & appropriate licences OPEN DATA
- Accessibility:
  - Finding the research data thanks to harvesting protocols based on standard metadata – *OLAC* *(Open Language Archives Community)*

OPEN LANGUAGE ARCHIVES

Hence in Mulce we decided to design corpora with these criteria in minds:

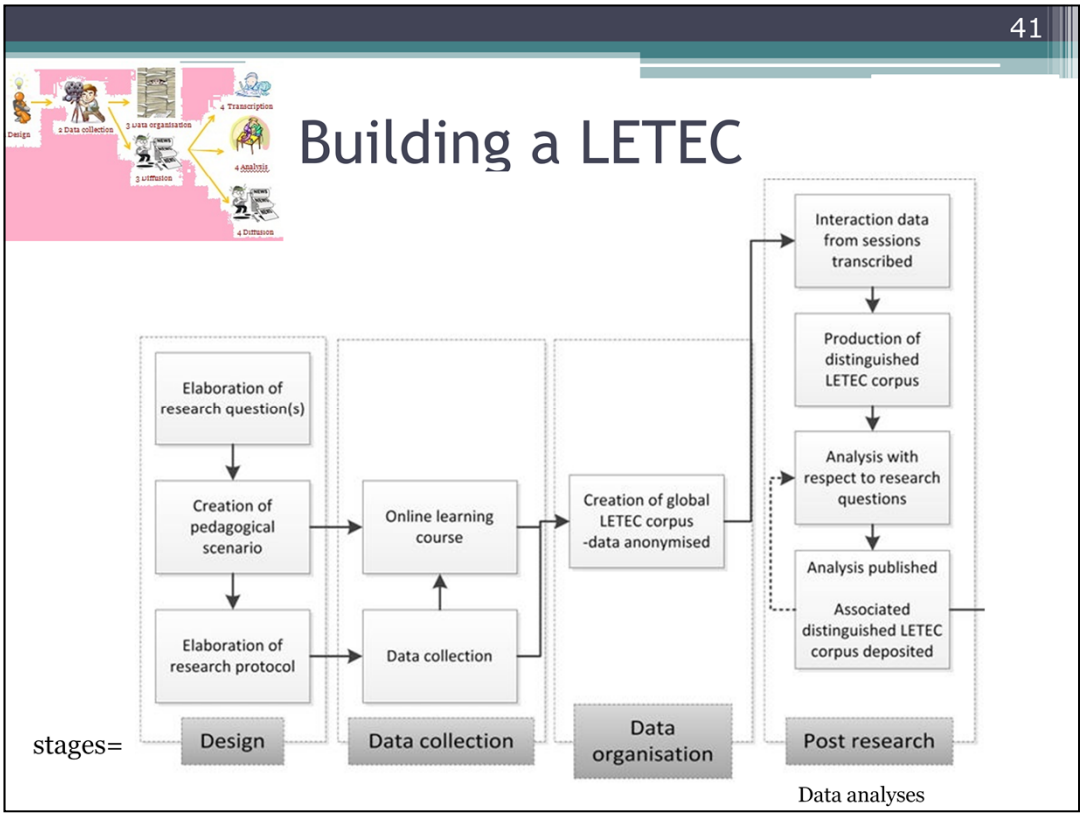
LETEC Components



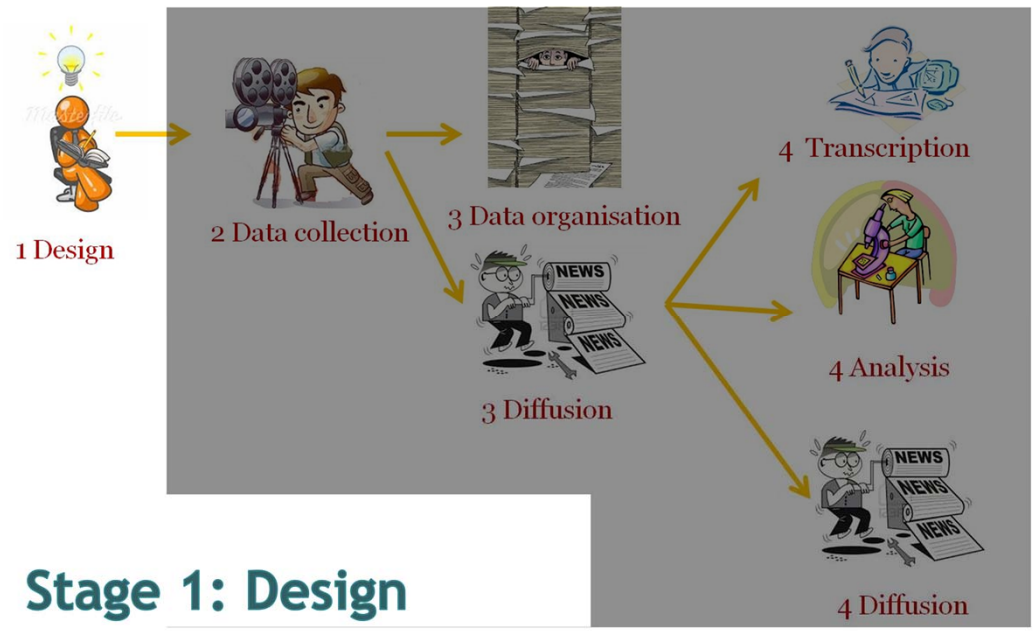
"A LETEC corpus collects in a **systematic and structured way** all the data from **interactions** which occur during a course which is **partially or entirely online**. These data are enriched by technical, pedagogical and scientific information as well as information about the participants and are organized to allow **contextualized analyses** to be performed." (Mulce-documentation, 2013)

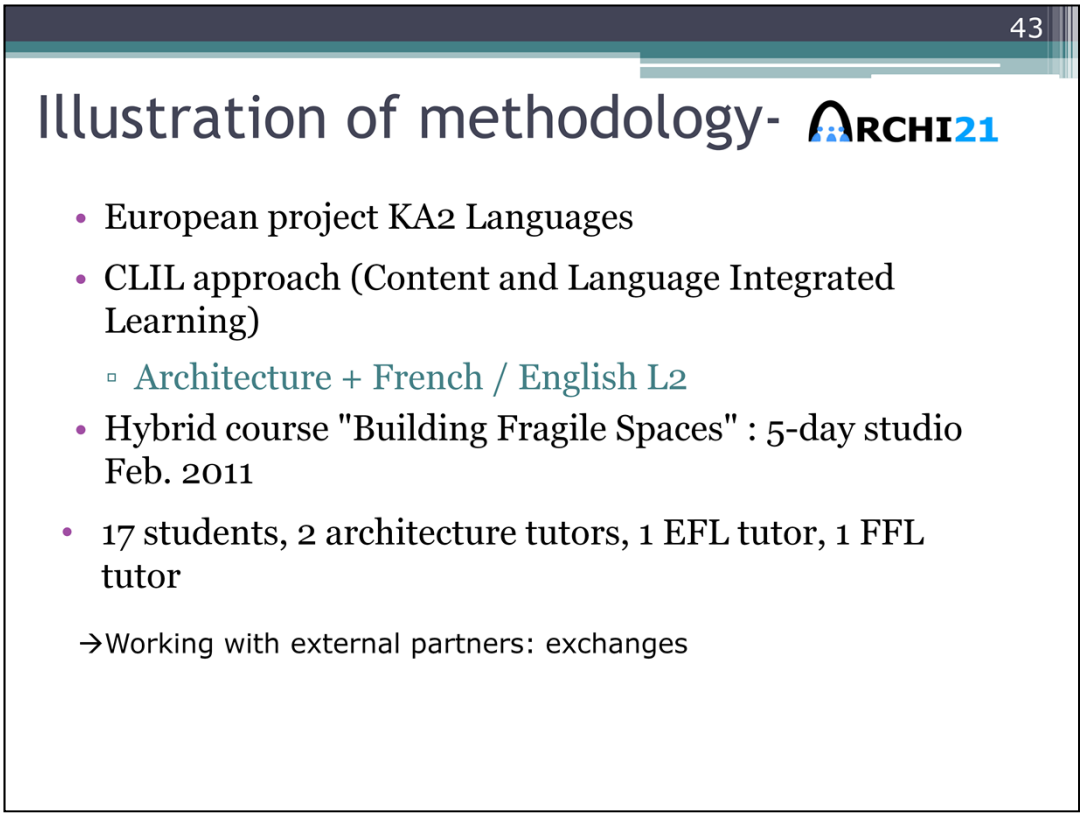
Here is the definition of what we mean by a LETEC corpus.

A learning and teaching corpus is made of several parts. The research protocol and the pedagogical scenario describe the context of the learning situation. The technical term "instantiation" (coming from the IMS consortium) refers to interactions and participants' productions collected during the learning situations. We also assemble forms and licences related to ethics and right. Forthcoming analysis may be attached to the first version of the corpus or come later on.



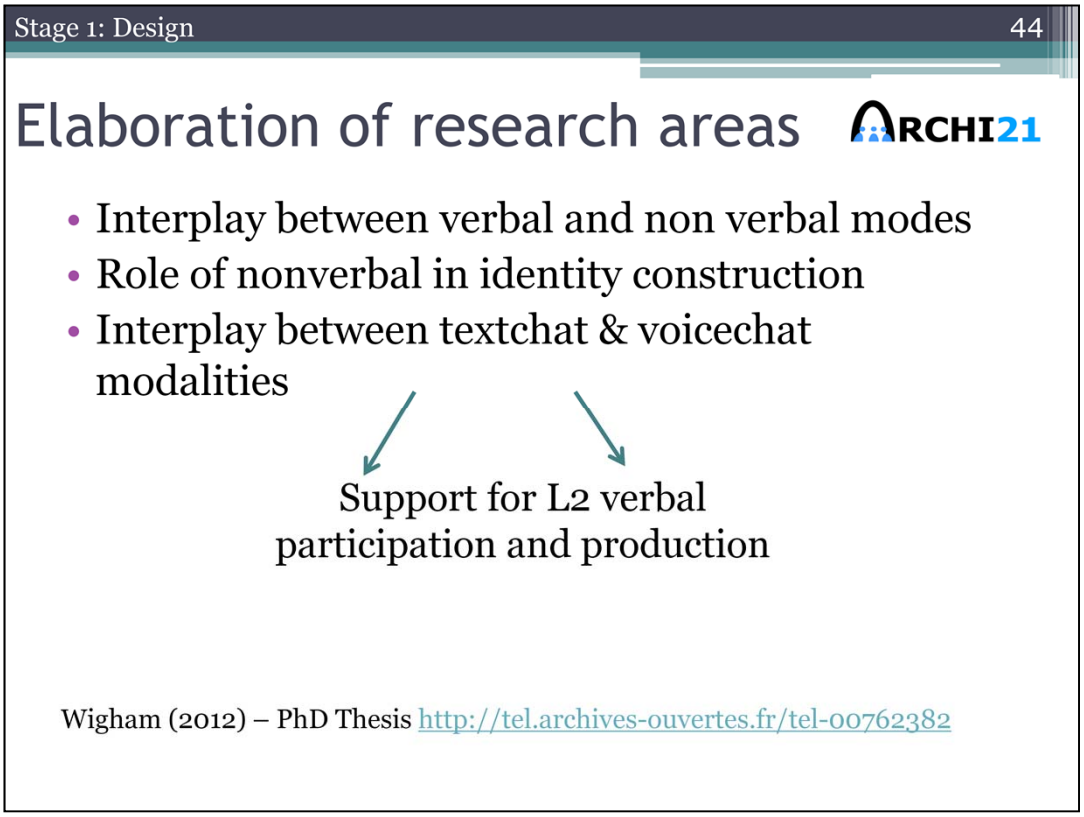
How can a LETEC be build? Here is a schema detailing the process. We presented it at WorldCALL. I will quiclky skip into it just for the sake of understanding the rest of my presentation.


1 Design
2 Data collection
3 Data organisation
3 Diffusion
4 Transcription
4 Analysis
4 Diffusion
NEWS
NEWS
NEWS
Stage 1: Design



To contextualized things let me choose one of our latest project, Archi21.

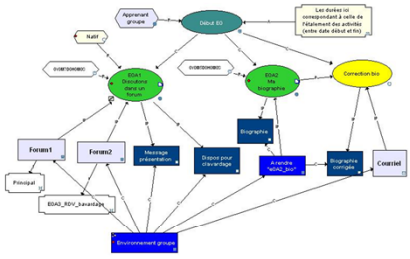
We designed a CLIL scenario jointly with teachers of architecture and language teachers. Learners had an intensive course in order to develop an architectural project, simultaneously F2F and online in Second Life. Language teachers were only at a distance. We had 4 groups of learners either in French or English as a foreign language.

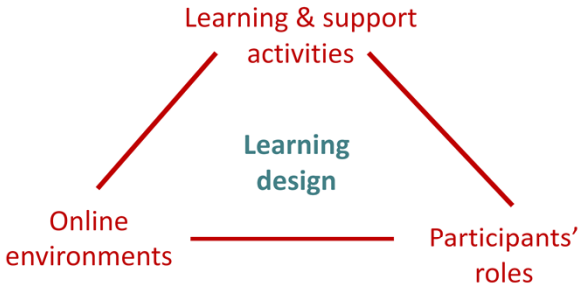


During this first stage, the design stage, research questions are fixed, here mainly:
- relationships beween verbal and non vernal modes (by “verbal” please understand it as being the antonym of non-verbal. No exclusive relation to speech)

- And Interplay between textchat & voicechat modalities

Pedagogical Design





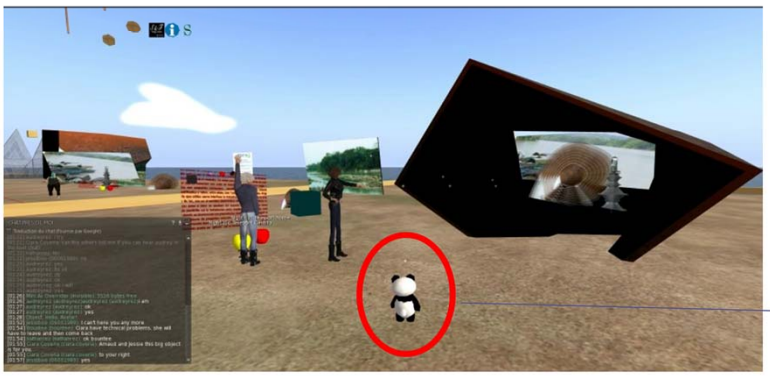
- Macro-task– collaboratively elaborate a model in a synthetic world (*Second Life*) as a response to an architectural problem brief
- Architectural studio, hybrid CLIL approach
- 4 workgroups

The design process of a learning scenario encompasses the description of the 3 main elements : online environments, learning activities and participant's roles.

Generally we present our learning design in a pretty formal way (see the graph on the left) in order to let people clearly understand it and to relate every step to pedagogical documents such as the guidelines and resources given to the learners. But there is no obligation for using such description format. A corpus compiler can just describes the pedagogical design as a simple text.

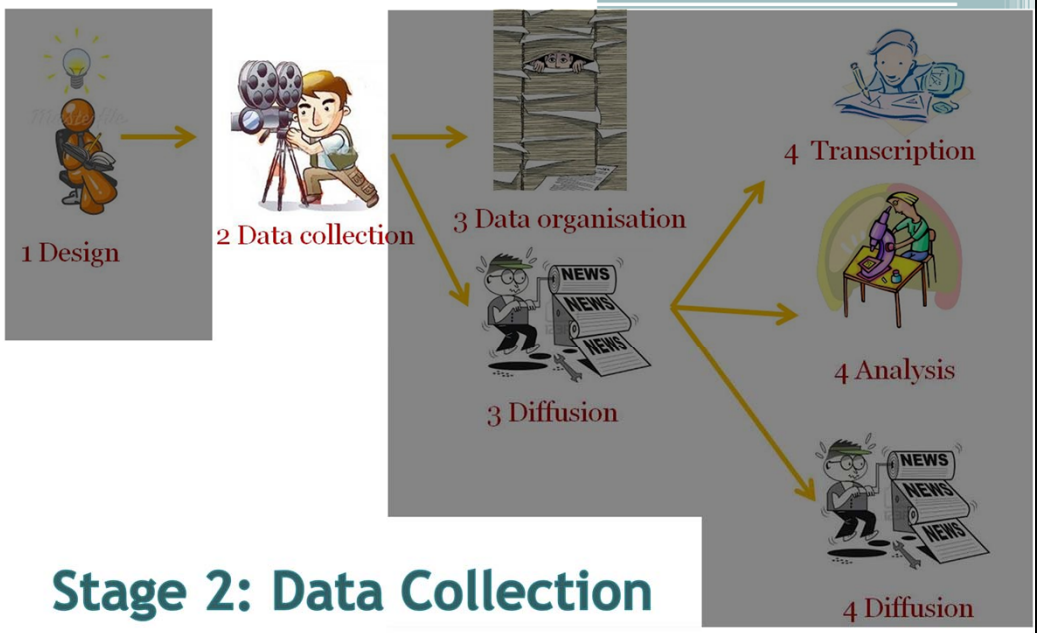
Research protocol

- Research protocol design
  - Protocol for data collection
  - Researchers' roles
  - Timetable of research activities

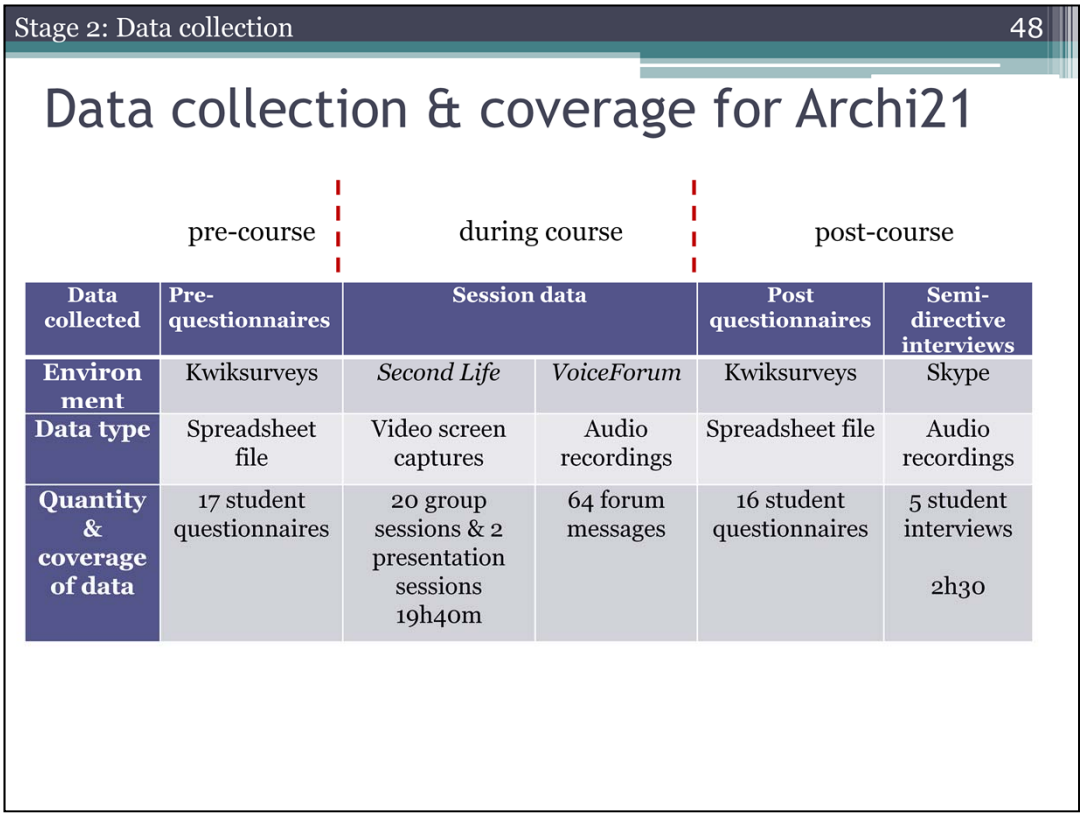


Wigham & Chanier, 2013 *ReCALL*

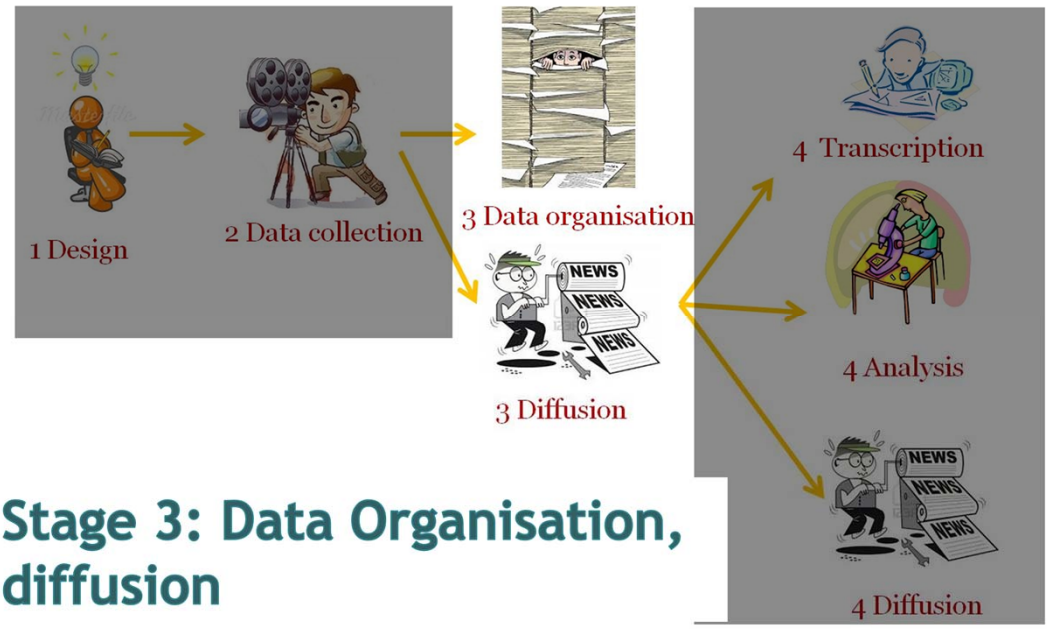
The design of the research protocol includes definition of researchers' role besides the teacher's one, the protocol for data collection, for questionnaires, ethics agreement, etc.



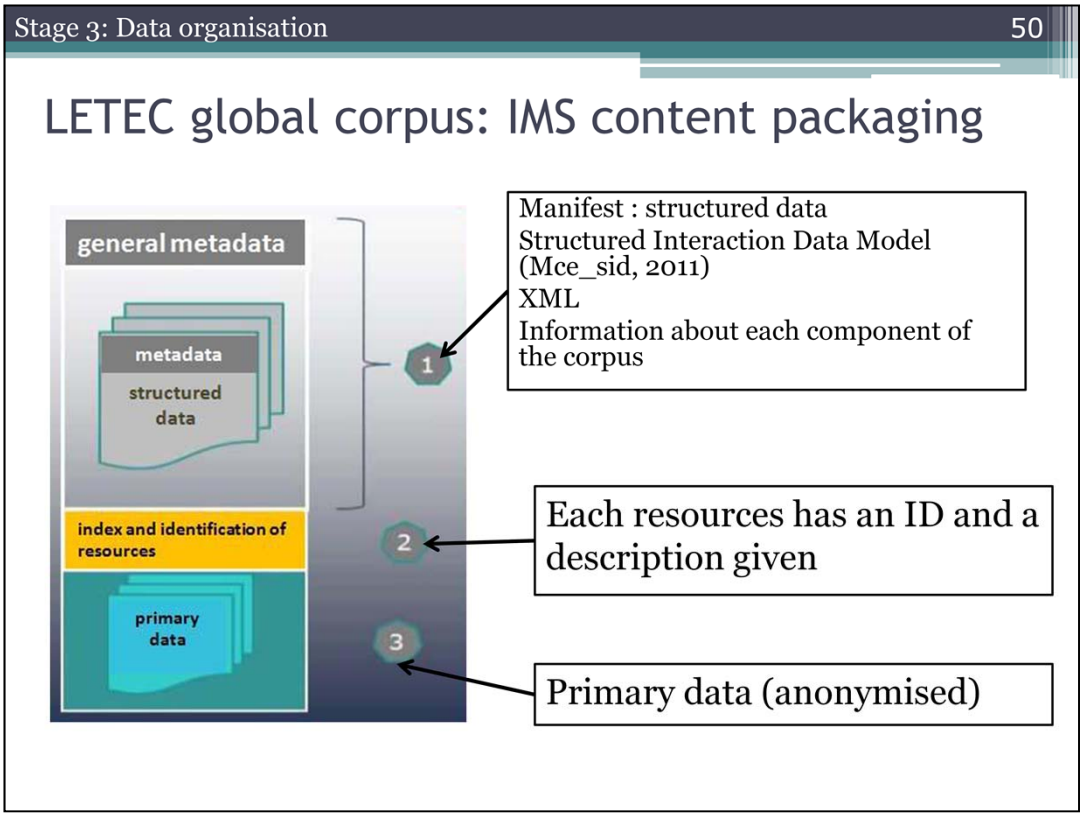
Stage 2: Data Collection



Here is an overview of data collected during the Archi21 course: pre and post questionaires, 20hours of videoscreen captures, post-interviews, etc.



Stage 3: Data Organisation, diffusion



In order to organize data, we use the IMS-CP format (a standard coming from the IMS consortium already mentioned. This international consortium which gathers academic institutions and companies is concerned with establishing interoperability for learning systems and learning content) , which I have no time to develop here.

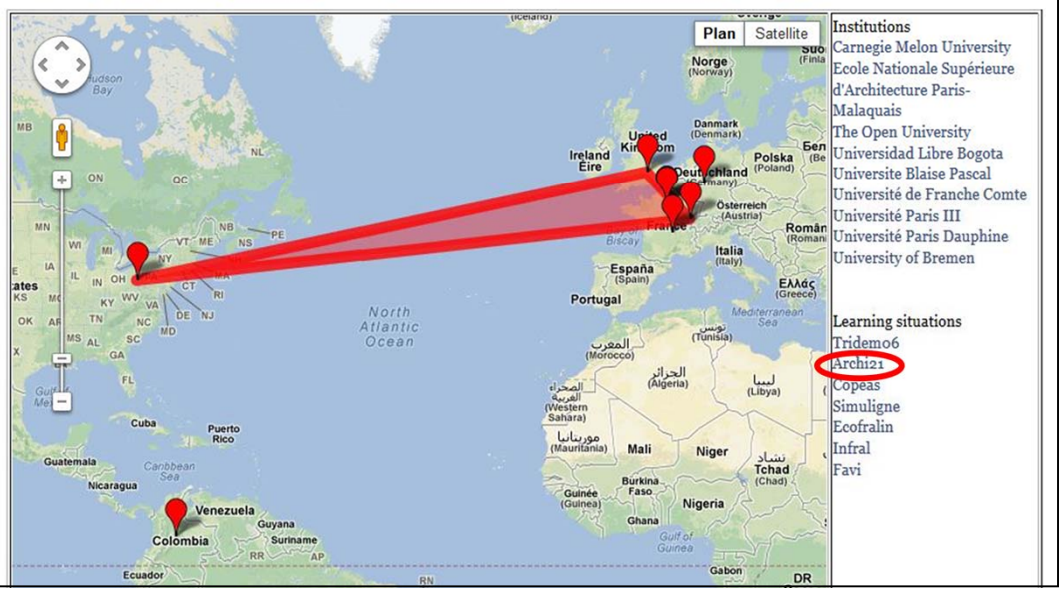
The bottom part of the corpus assemble the primary data (after they have been anonymised). In the second part, a set of IDs make the link between the top description and the corresponding files of the primary data.

The top part, called the "manifest" (another technical term) is made of one XML file and give information about each component of the corpus: metadata, technological environments used in the course, bio information and IDS about participants (teachers, learners, groups).
Then in a structure called workspaces are the interactions, links between what participants did with respect to the pedagogical

scenario.

Corpus deposit

- *Mulce* corpus repository : http://repository.mulce.org



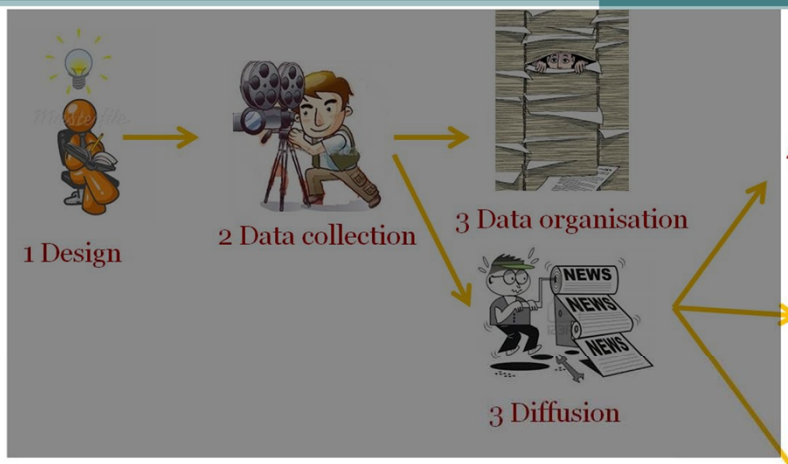
All these 3 components are assembled in an archive which is deposited into the corpus databank, “Mulce repository”.

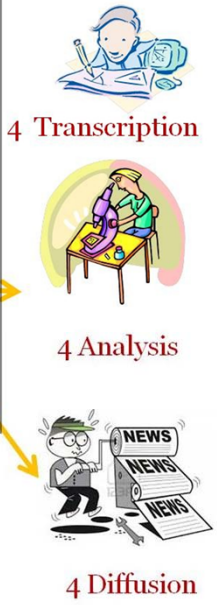
Here is part of the interface with the locations of the different experiments already mentioned. It is one way to access the archives / corpora.

Corpus diffusion

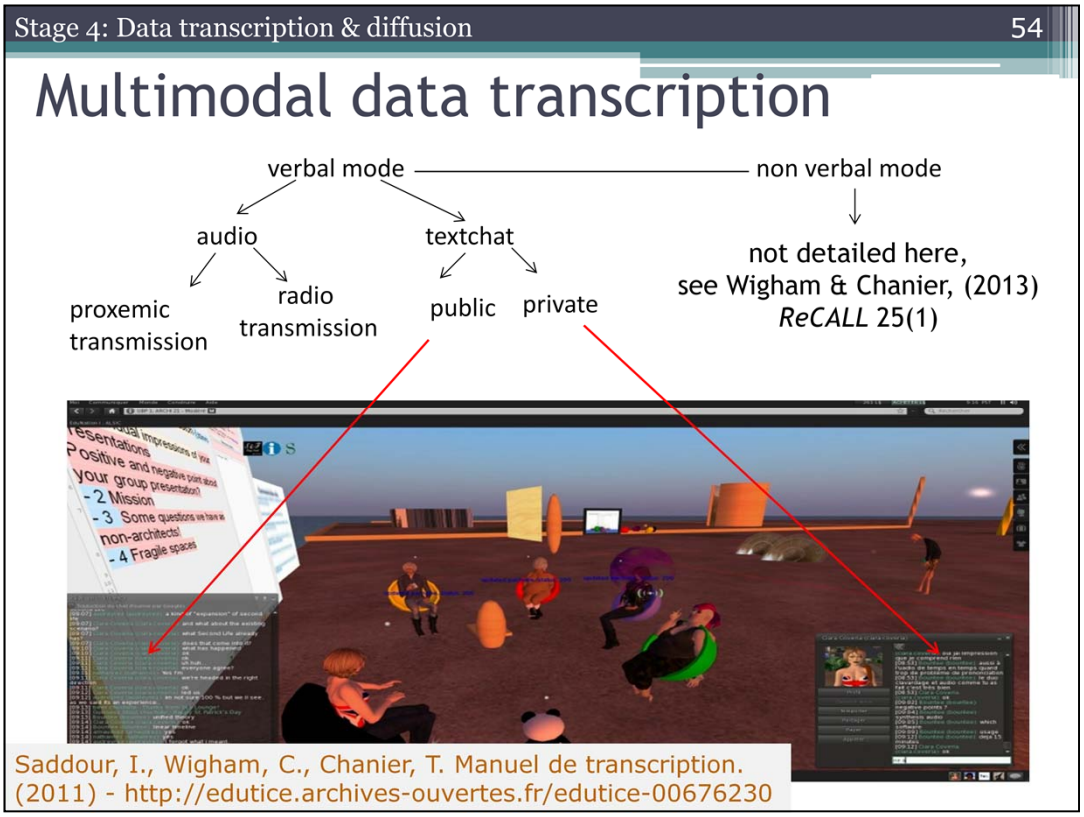
- Description of corpus; interface to browse structure; zip file to download

Another way of selecting and downloading corpora is offered through this second interface which details corpora criteria such as, learning situations, technological environments used, languages, types of interactions, pedagogical approaches, etc.

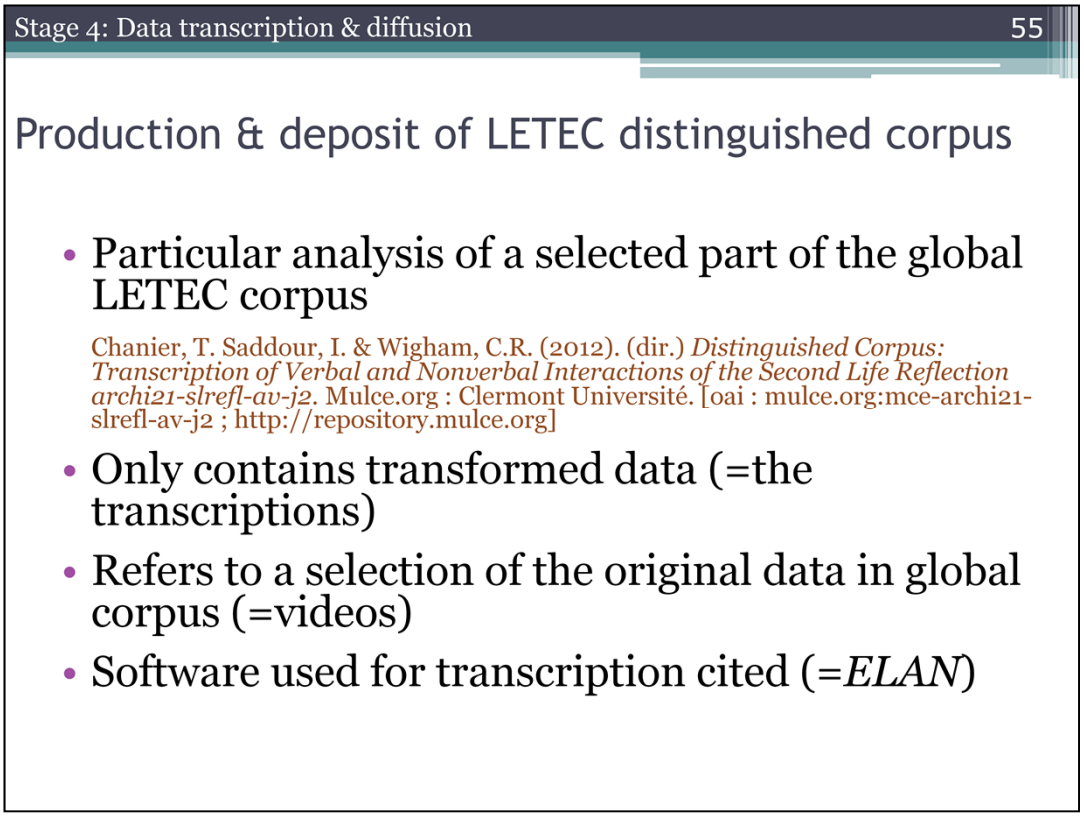




Stage 4: Transcription, analyses, publications



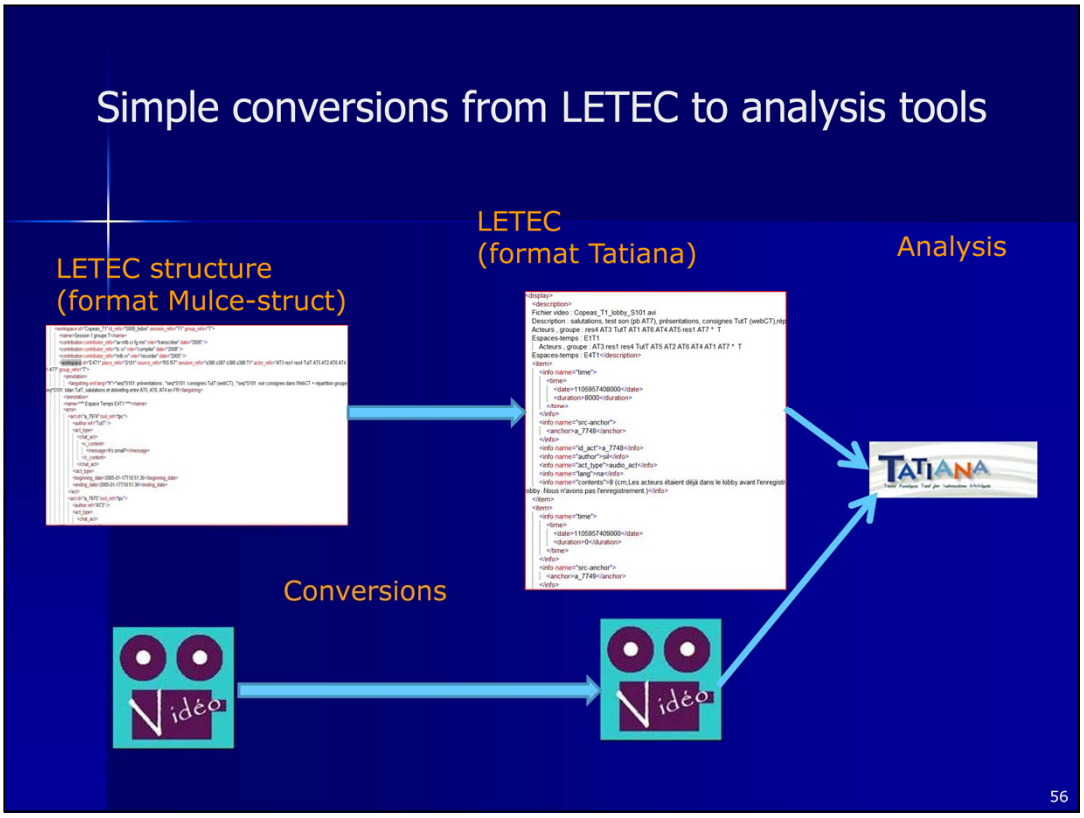
We started transcribing multimodal data in 2005 thanks to Lyceum, the Open University environment used in the Copéas experiment. Here is a simplified view of the kinds of interactions we transcribe in Second Life (it is detailed in one of our recent paper in ReCALL).
Whilst teaching and researching in various environments, we elaborated our transcription methodology. The latest version of our manual for transcription is online and, as all our publication, open access.



Once you achieve your transcription and analysis, you compile them into what we called a distinguished corpus (the one before being called the global corpus) and you deposit this second corpus.

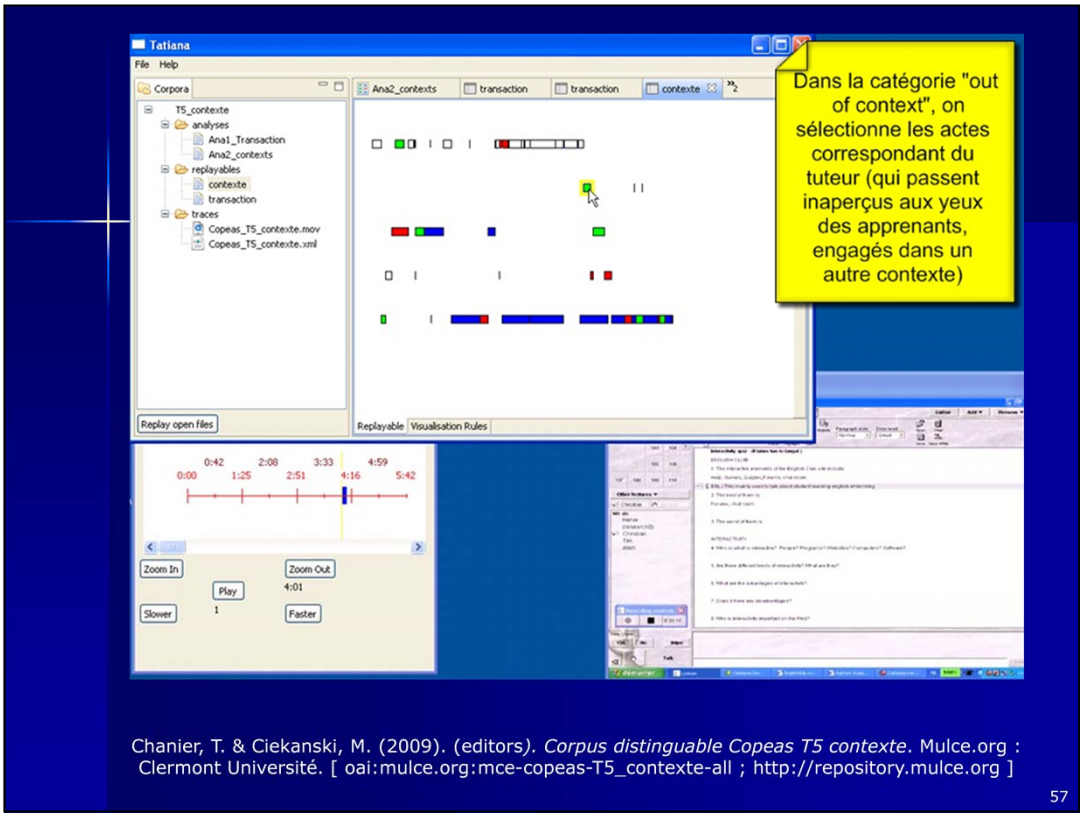
As you can see here every corpus receive its own reference on the repository.

The new corpus only contains transformed data. It gives links to data already described into the global corpus, and add description on the tools used during the analysis and the transcription step.



Some colleagues often wonder why spending so much time on organizing data. Well research is not a one shot process. Once you have your learning situation and your data, you may want to study it from different perspectives. News ideas may come and new analysis tools help you build these new ideas, it is an interactive process.

For example here, on the left you have transcriptions coming from a Copeas corpus, trancriptions in our XML format, linked to a video capture in a given format. We have been interested in using the Tatiana software. It has been a straightforward process to transform Mulce XML format to Tatiana XML one and convert videos formats..



As an example, you can see an extract of a session in Lyceum, displayed in Tatiana : colors refers to modalities (audio, textchat, word processor) and layers to participants. 3 learners are collaboratively writing into a document, whereas the tutor tries several times intervening but is completely ignore by the learners.

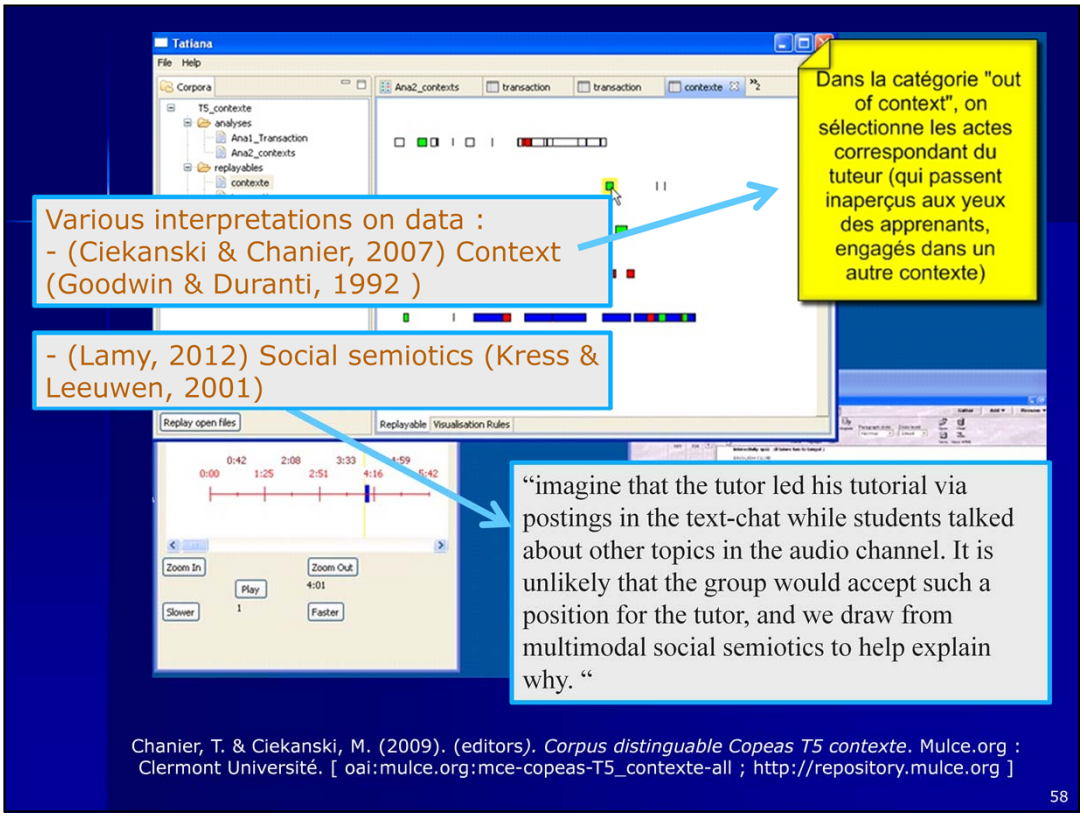
Maud Ciekanski gave an interpretation to this surprising phenomena through an analysis of the Context. She used the Goodwin & Duranti 92 model to explain why the tutor was out of context. We published this analysis as a LETEC corpus.

Funnily I recently discovered in a paper published by Lamy in 2012, this quotation.

It is a general paper were she explains why Social semiotics can be a good theoretical framework to study multimodality.

As far as I understood, Marie-Noelle was not referring to a specific example.

We can feel here how it could be interesting, around pieces of data, to raises different theoretical issues, or explanations.



As an example, you can see an extract of a session in Lyceum, displayed in Tatiana : colors refers to modalities (audio, textchat, word processor) and layers to participants. 3 learners are collaboratively writing into a document, whereas the tutor tries several times intervening but is completely ignore by the learners.

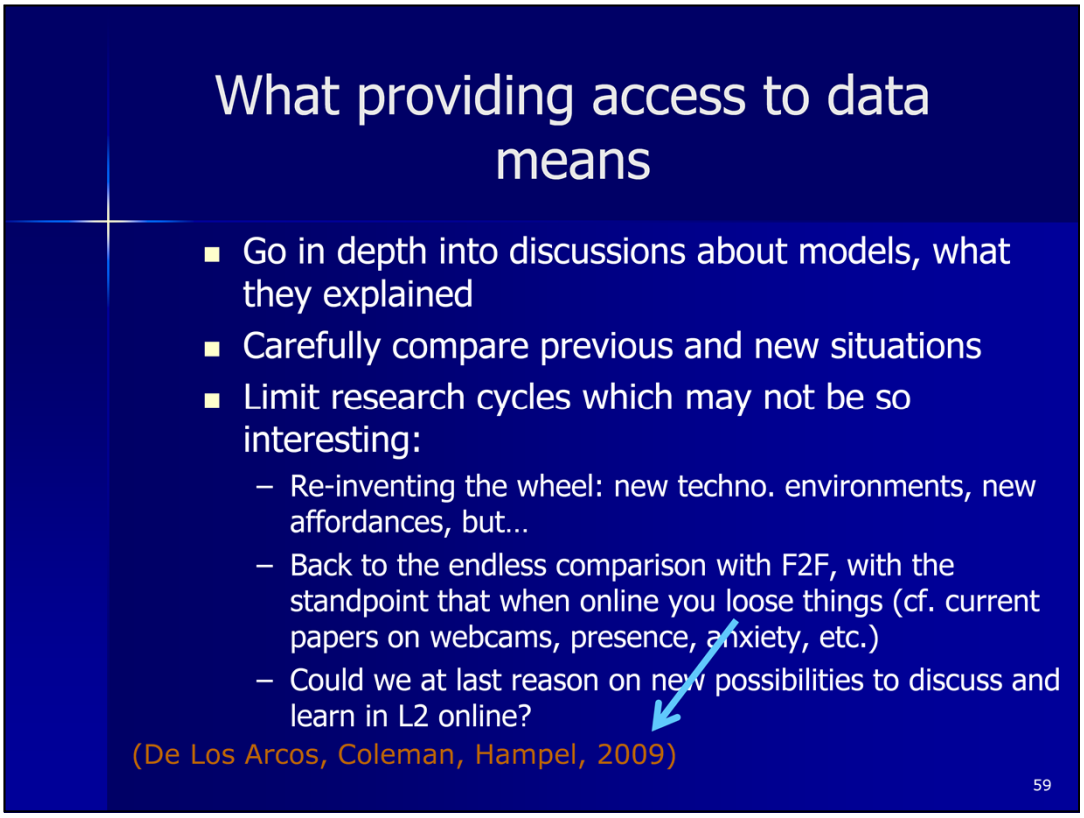
Maud Ciekanski gave an interpretation to this surprising phenomena through an analysis of the Context. She used the Goodwin & Duranti 92 model to explain why the tutor was out of context. We published this analysis as a LETEC corpus.

Funnily I recently discovered in a paper published by Lamy in 2012, this quotation.

It is a general paper were she explains why Social semiotics can be a good theoretical framework to study multimodality.

As far as I understood, Marie-Noelle was not referring to a specific example.

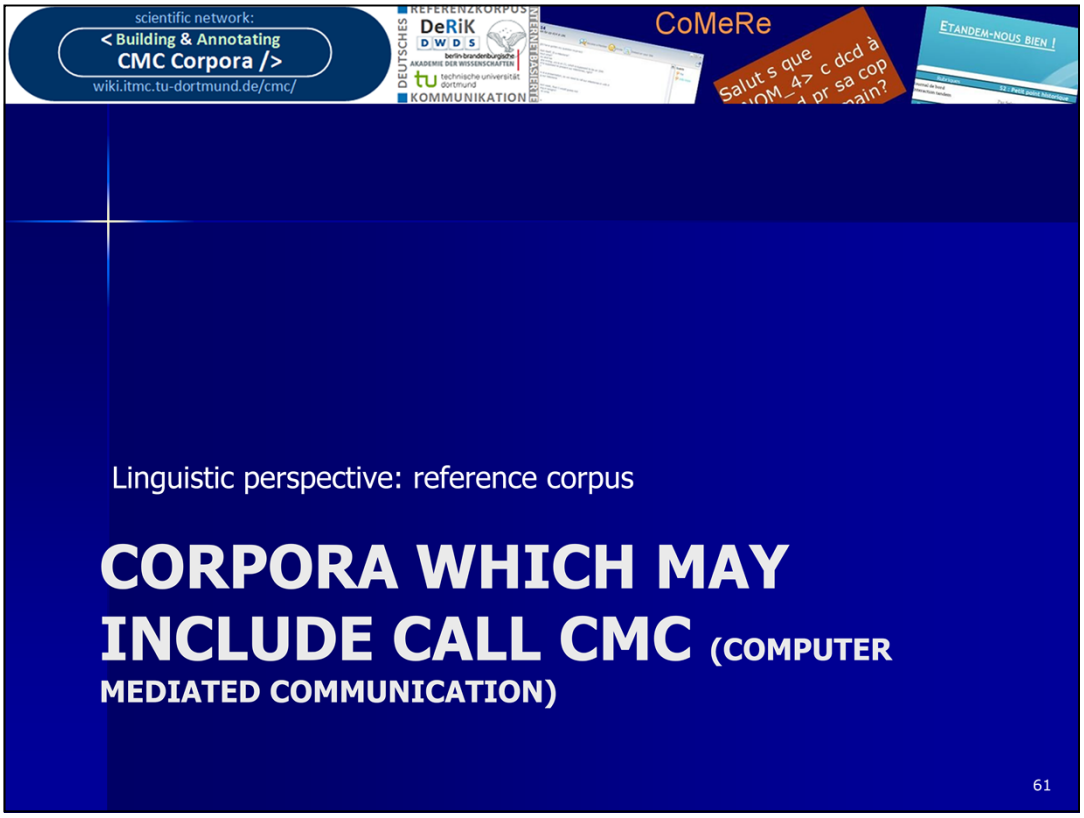
We can feel here how it could be interesting, around pieces of data, to raises different theoretical issues, or explanations.



What can we gain when giving access to research data?

De Los Arcos paper demonstrate for example that anxiety is not present in the audiographic environment and consequently this type environment needs to be studied as specific topic.

Organising and publishing data may be useful to CALL research. But we can extend our perspective and see whether it may be of interest more generally in linguistics or for pedagogical motivations.



Let us firstly consider the linguistic perspective, and keep in mind that we all here in CALL have got very rich CMC data.

Reference corpora of different languages

- Corpus in German, DWDS
  Digitales Wörterbuch der deutschen Sprache, http://www.dwds.de/
- Corpus in Flemish / Dutch, SoNaR
  ***S**TEVIN **N**ederlandstalig **R**eferentiecorpus*
- Corpus in French (in progress)
- Common aims:
  - Billions of tokens, 500 M structured & annotated (POS), access for linguistic research
  - Extension to Internet communication

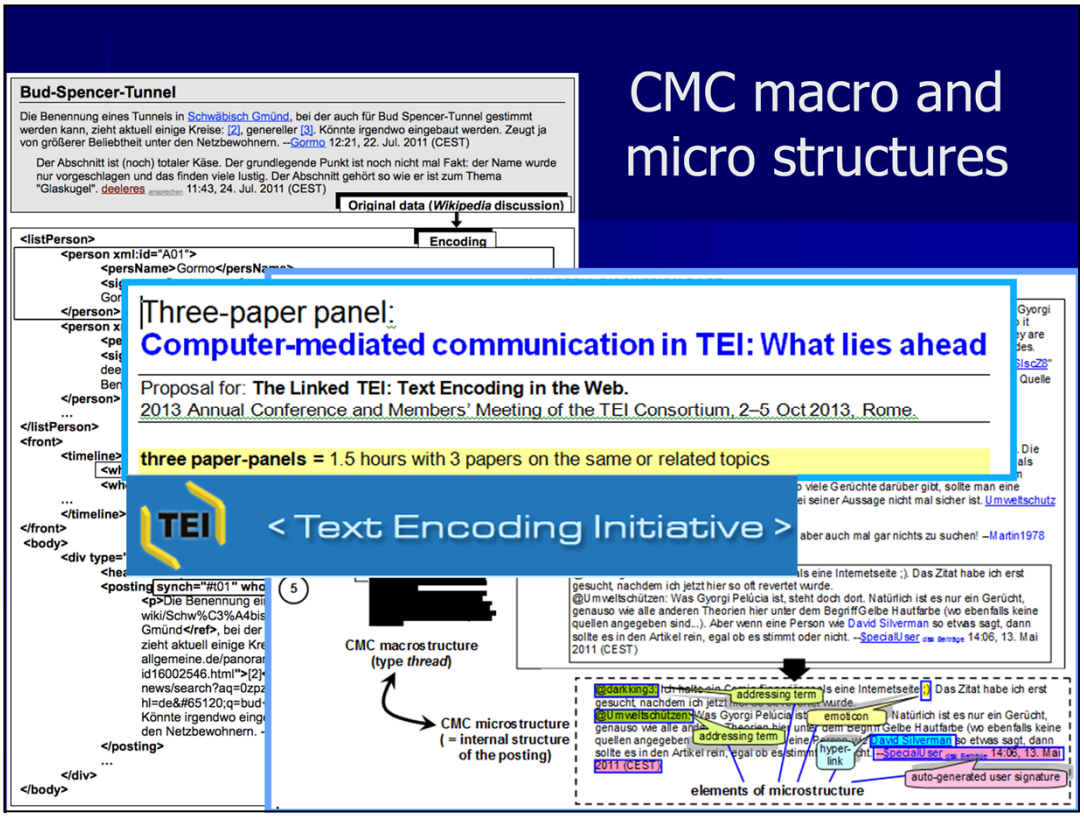
62

After the historical project on the English language which lead to the creation of the BNC corpus (British National Corpus), during the last 10 years linguists started to develop reference corpora of other European languages started :

- There is a reference corpus of German
- A second one for Flemish Dutch
- Another reference corpus for French is in progress.

Their common features are :

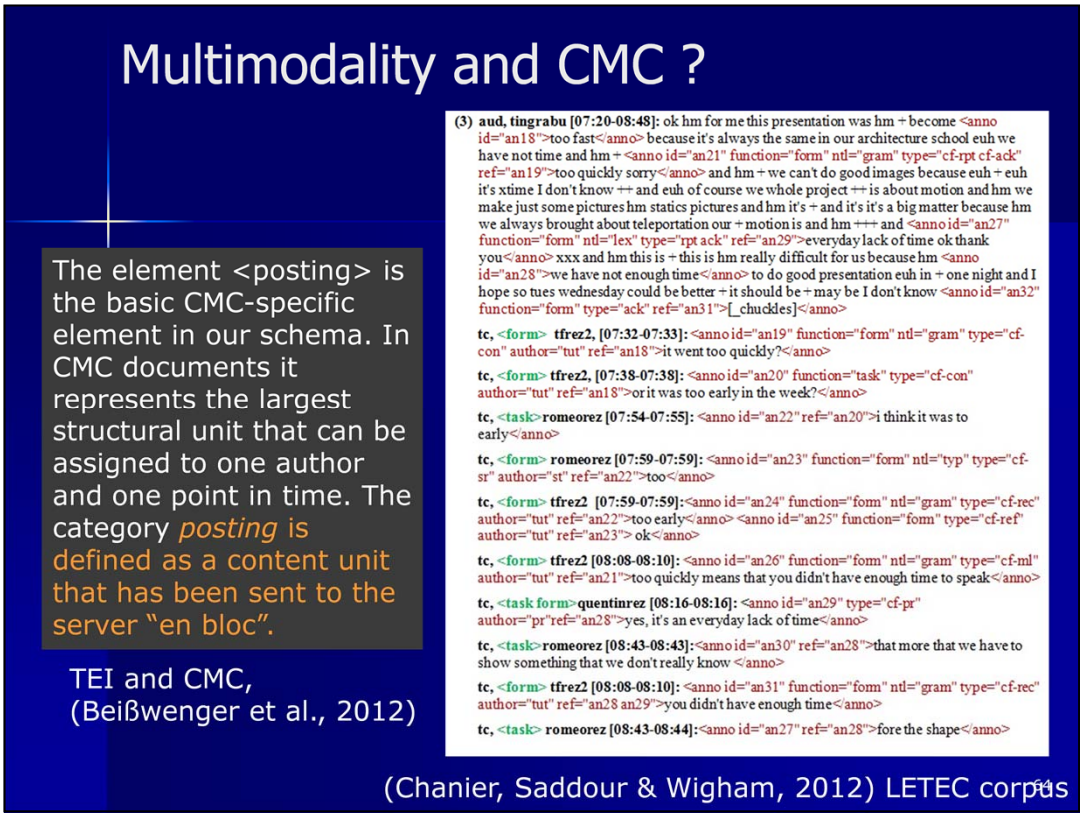
- Large couverture, Billions of tokens, 500 M structured & annotated (POS),
- Provide access for linguistic research
- They seek to develop extensions to Internet communication



When colleagues in Linguistics build corpora they systematically structure their data.

Hence when considering CMC data, one of the first thing to do is study their specific micro and macro structures (here Wikipedia forums).

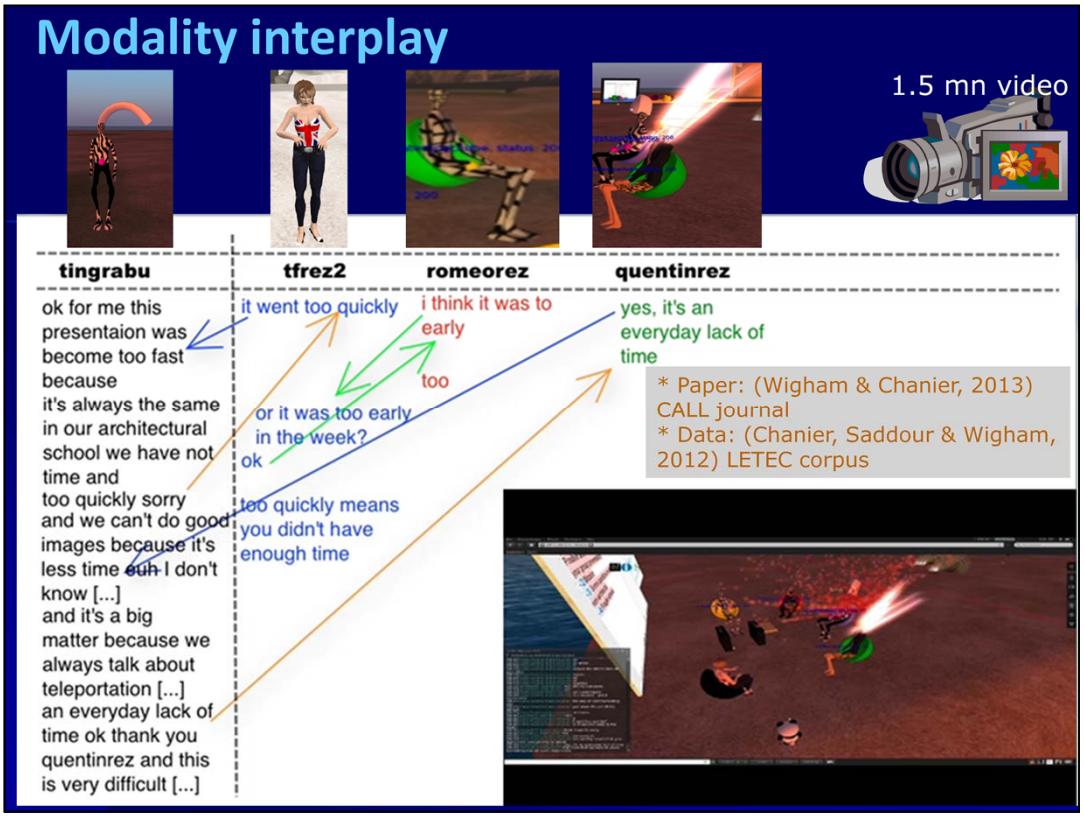
Very often they use the TEI (Text Encoding Initiative) , a standard previously designed for text, then extend to speech. They now want to propose another CMC extension to the TEI.



My German linguist colleague, Michael Beisswenger, after studying discussion forums and textchats recently characterized CMC structures.

He sais …

But in CALL, when we study multimodal CMC the "en bloc" nature does not apply anymore. Here is a transcription of a LETEC corpus which shows interplay between modalities. Since it s not very readable, let us see it differently.



In one of our paper, which will appear in the CALL journal, and the corresponding data are already online in Mulce, Ciara Wigham discusses the interplay between audio and textchat.

Here is an extract from Archi21. In the left column you have the transcription of the audio of one learner, who presents his feeling related to the on-going process of his architectural project. He is a French native and speaks in English as his L2. In the 3 other columns on the right, you find textchats turns coming from the tutor and two other learners belonging to the same architectural project group.

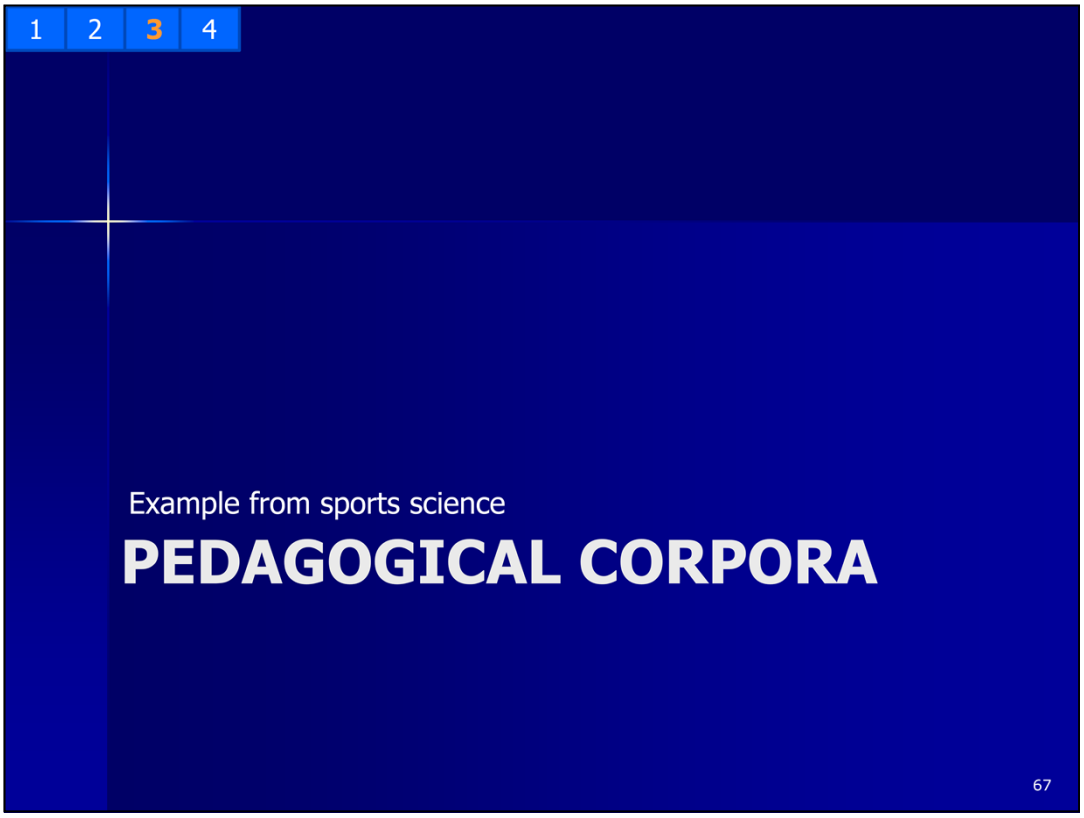
Let me show you a short video.

**** In this example of conversation doubling, the acts in the text chat respond to the voice chat (blue arrows) but equally acts in the voice chat respond to the text chat (orange arrows) and text chat acts respond to interaction in both voice chat and text chat modalities and prompt interaction in both modalities

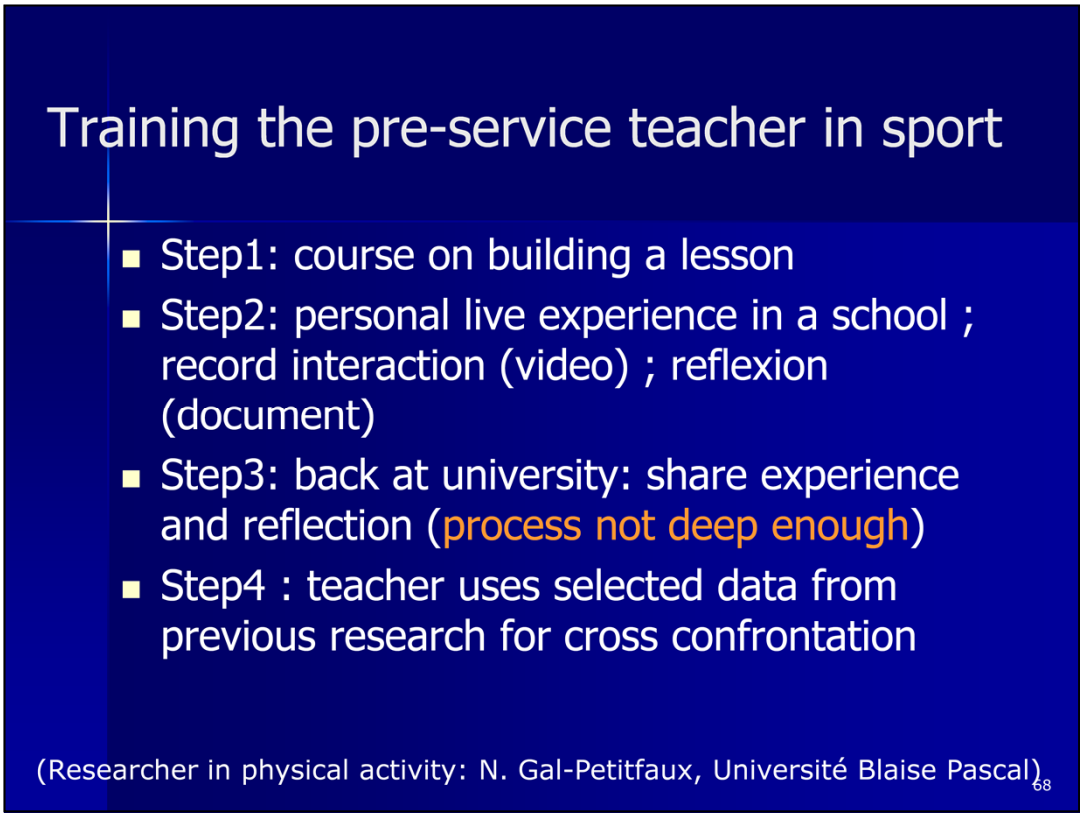
Thanks to the rich kind of data we find in CALL, we have been able to create with colleagues belonging to a dozen of different linguistic research labs the CoMeRe project.

CoMeRe is the French name given to CMC.

We aim at building a CMC corpus in French and to participate at the same time at a European level, with Michael Beisswenger, to the extension of the TEI to CMC.



Let us now consider pedagogical applications of LETEC corpora



The idea of pedagogical corpora stemmed out of discussions with colleagues who are simultaneously researchers in physical activity and teacher trainers in sport. Here is how they design new procedures to train pre-service sport teacher.

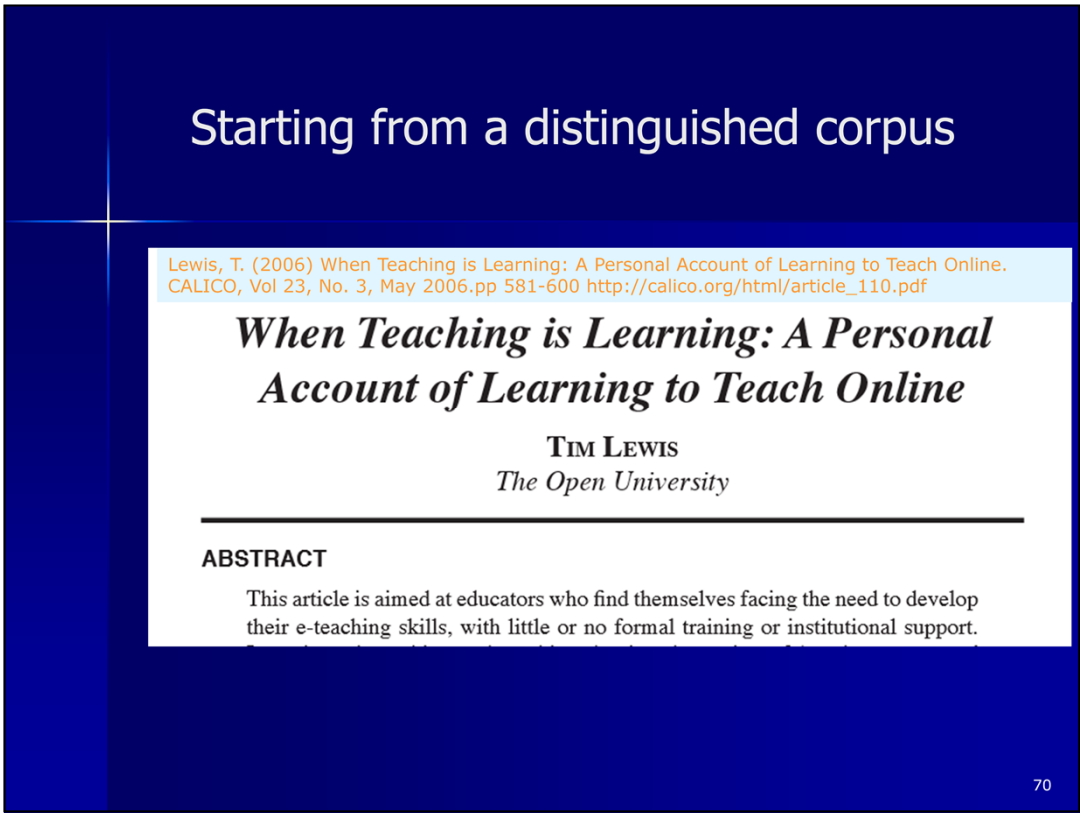
- step1 : the teacher trainer explains how a sport lesson should be designed

- Step 2 : students, organized by couples, have a live experience : one teaches a lesson in a school ; the second one records the lesson

- Step3 : students come back at the university, they share their experience, use their videos. But the reflection process is not deep enough

- hence step 4 : the teacher trainer uses selected data from previous research situation for cross confrontation.

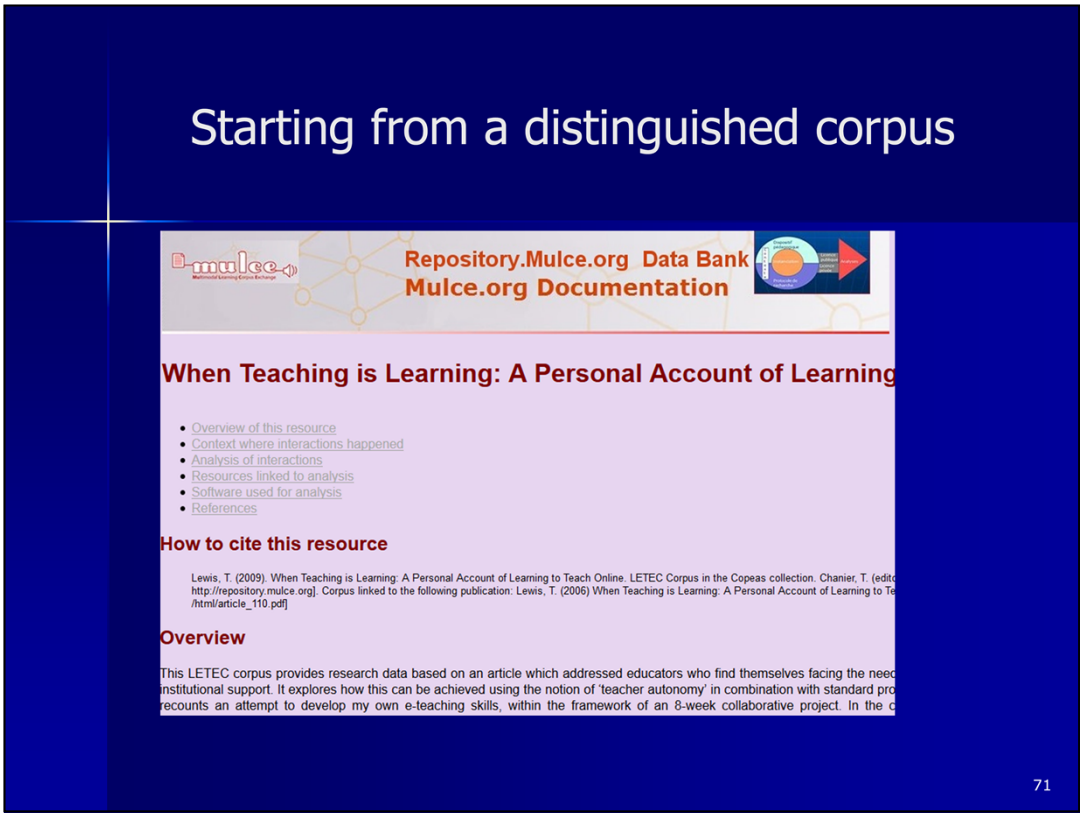

1 2 3 4
Authors : Ciara Wigham, Thierry Chanier
PEDAGOGICAL CORPORA CREATED OUT OF LETEC CORPORA
69



To understand the flavour of pedagogical corpora, let us consider a specific situation. In 2006, our colleague Tim Lewis, from the Open University, who have had his first experience as online tutor in a multimodal environment during the Copeas experiment, published afterwards a paper in the CALICO journal. We assembled in a LETEC corpus all the data he gave us, related to this paper : personal diary, videos interviews of learners. We added other data coming from learners reports, discussion forums. Tim was quite pessimistic about the nature of the collaborative process among the learners, and also between himself and the learners. But when you closely look at the data different perspectives appear.

Directly using a LETEC corpus in a training situation is not that easy.

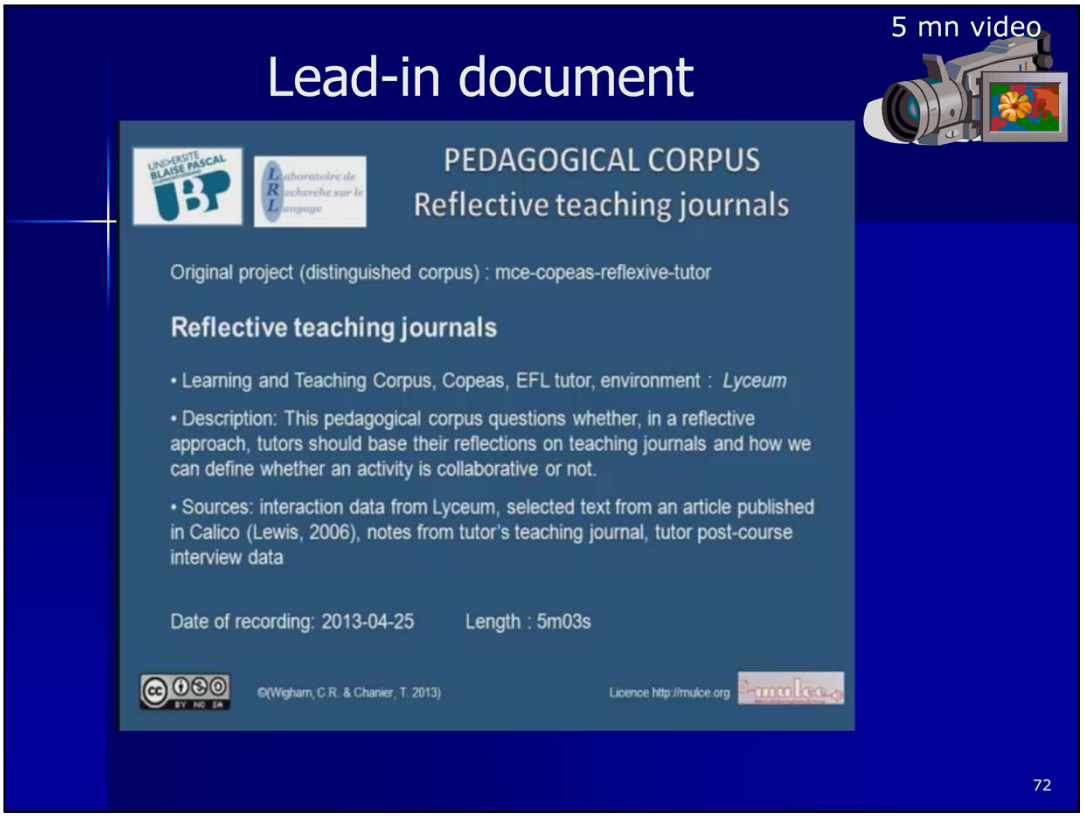
Hence we extracted data, and with Ciara Wigham, we imagine a pedagogical corpus.



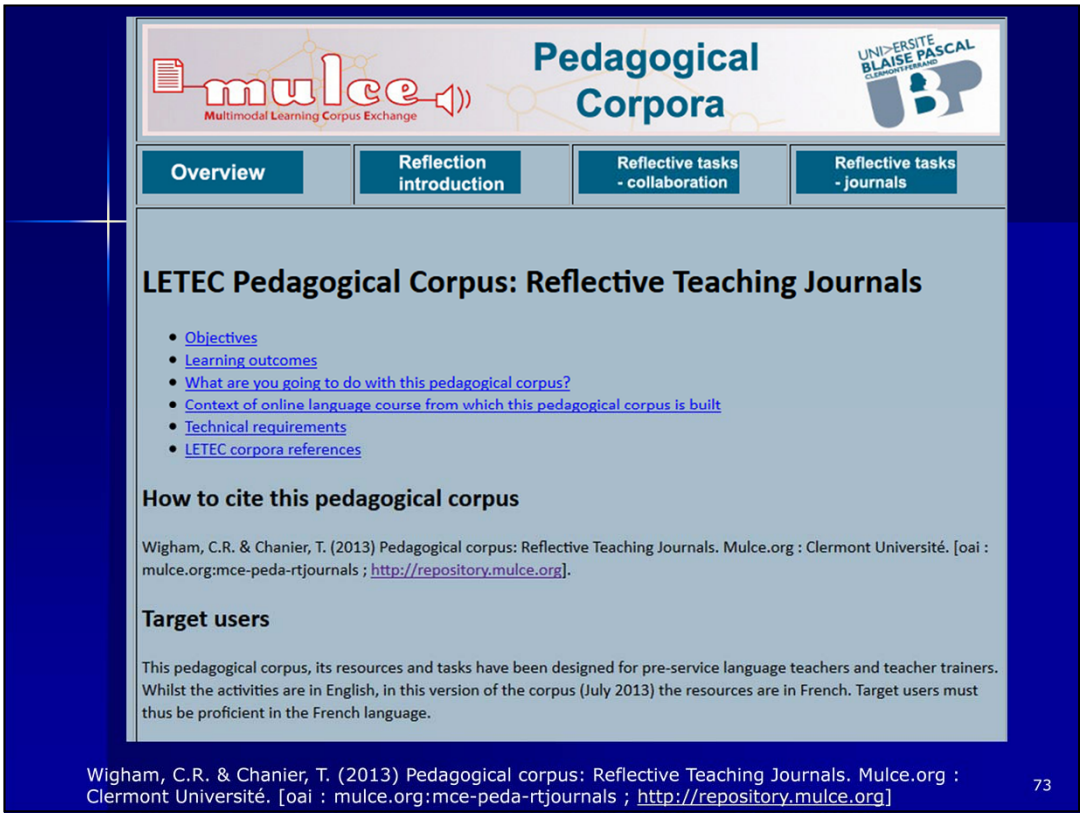
To understand the flavour of pedagogical corpora, let us consider a specific situation. In 2006, our colleague Tim Lewis, from the Open University, who have had his first experience as online tutor in a multimodal environment during the Copeas experiment, published afterwards a paper in the CALICO journal. We assembled in a LETEC corpus all the data he gave us, related to this paper : personal diary, videos interviews of learners. We added other data coming from learners reports, discussion forums. Tim was quite pessimistic about the nature of the collaborative process among the learners, and also between himself and the learners. But when you closely look at the data different perspectives appear.

Directly using a LETEC corpus in a training situation is not that easy.

Hence we extracted data, and with Ciara Wigham, we imagine a pedagogical corpus.



Here is an extract of a video which has been created as a lead-in document for the new pedagogical corpus.



The pedagogical corpus can be downloaded out of Mulce repository.

It uses data from the Copeas experiment and offer new activities around the perception of collaborative process as a language tutor.

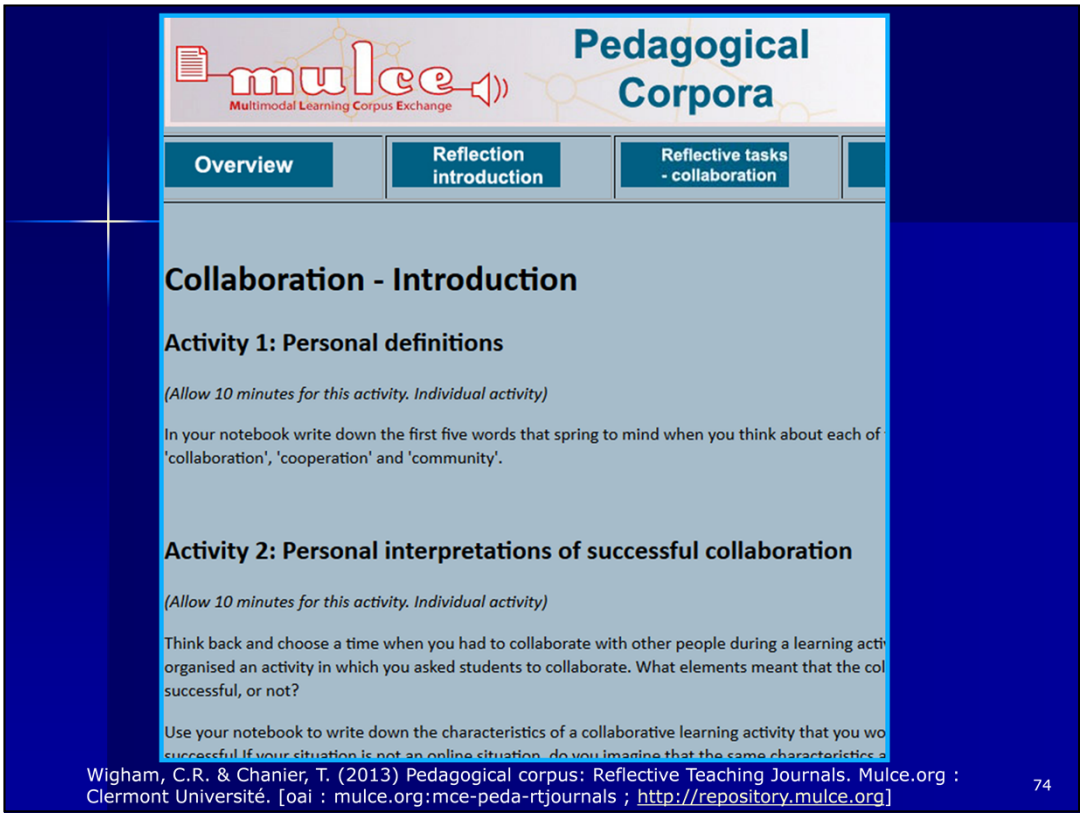
It aims at

- identify language tutors' and students' differing views of successful collaboration

- summarize the characteristics of successful collaboration and produce a list of implications for practice

- appraise the advantages of keeping a teaching journal

- compare and contrast reflections from a teaching journal with naturally occurring data (interaction tracks) and researcher-provoked data (student feedback) to analyse whether teachers should base reflections about teaching practice solely on journal entries and personal reactions



The pedagogical corpus can be downloaded out of Mulce repository.

It uses data from the Copeas experiment and offer new activities around the perception of collaborative process as a language tutor.

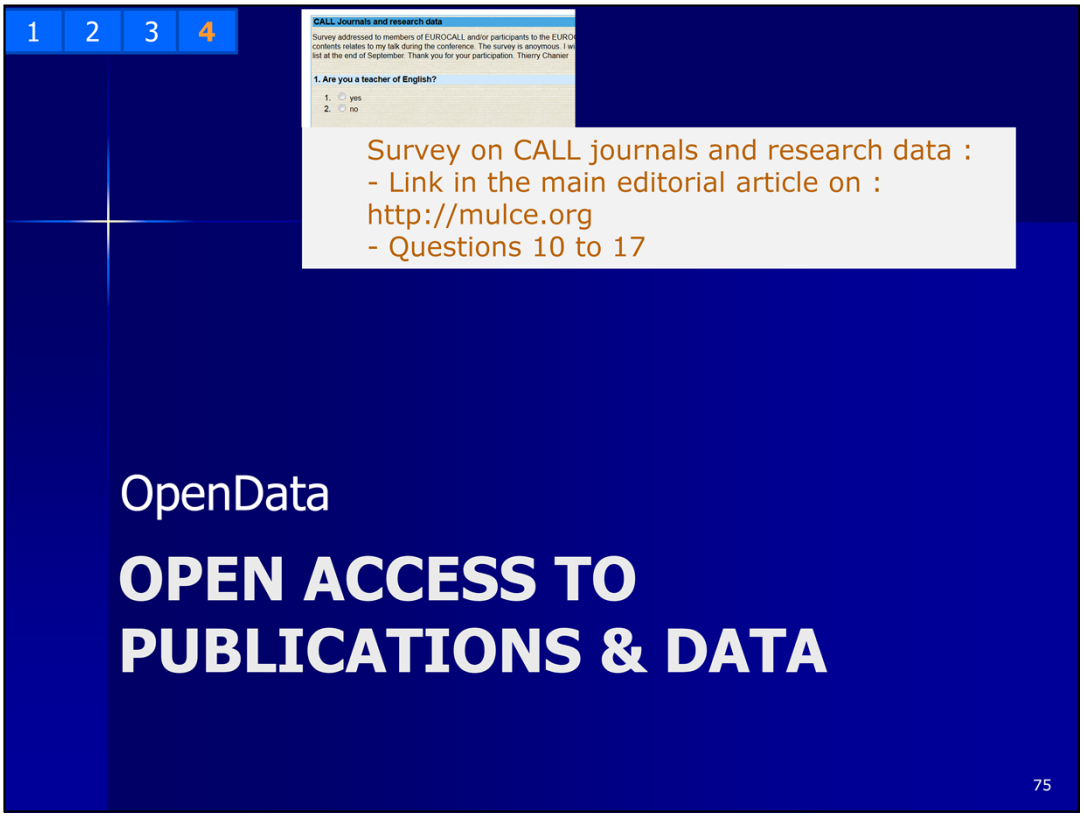
It aims at

- identify language tutors' and students' differing views of successful collaboration

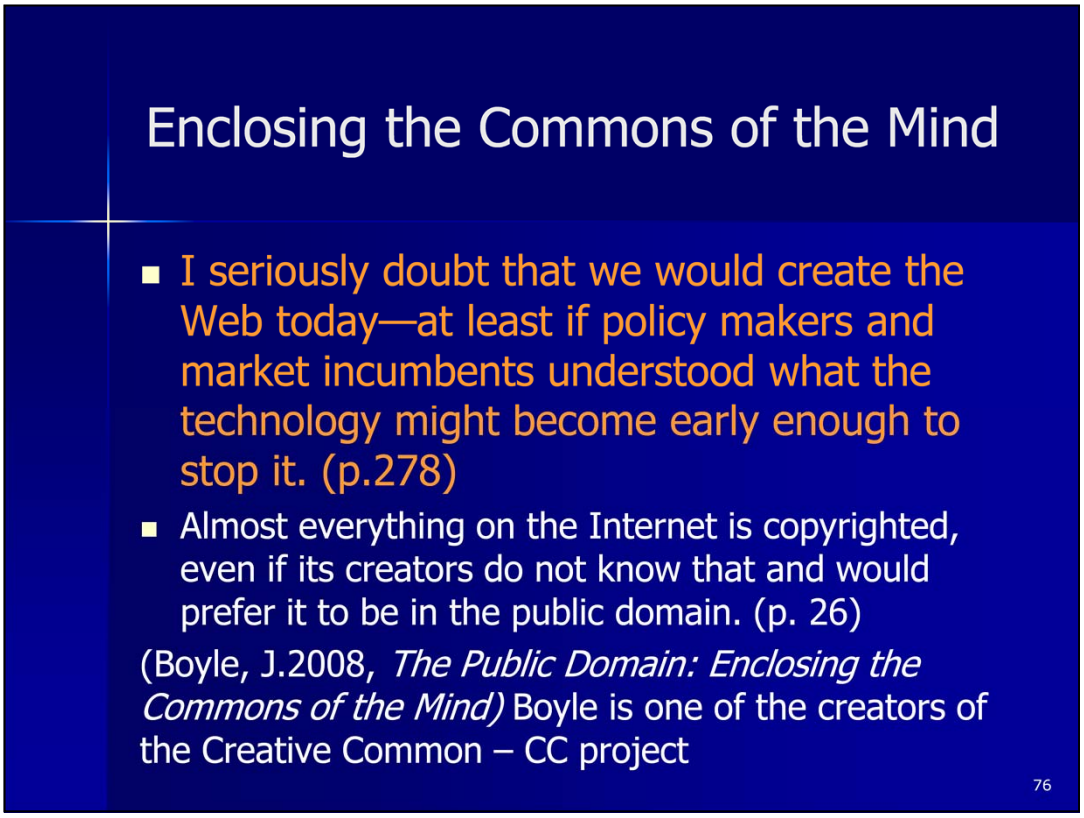
- summarize the characteristics of successful collaboration and produce a list of implications for practice

- appraise the advantages of keeping a teaching journal

- compare and contrast reflections from a teaching journal with naturally occurring data (interaction tracks) and researcher-provoked data (student feedback) to analyse whether teachers should base reflections about teaching practice solely on journal entries and personal reactions



4th section , last subject



Let us start with a warning coming from James Boyle, one of creator of Creative Common project. Many people would like to enclose the Commons of the Mind as he says.

I seriously …

And he gave a second warning about the necessity of paying attention to licences even when looking at public domain issues.

Let us briefly look at open access to publications, subject I already detailed in LLT in 2007..

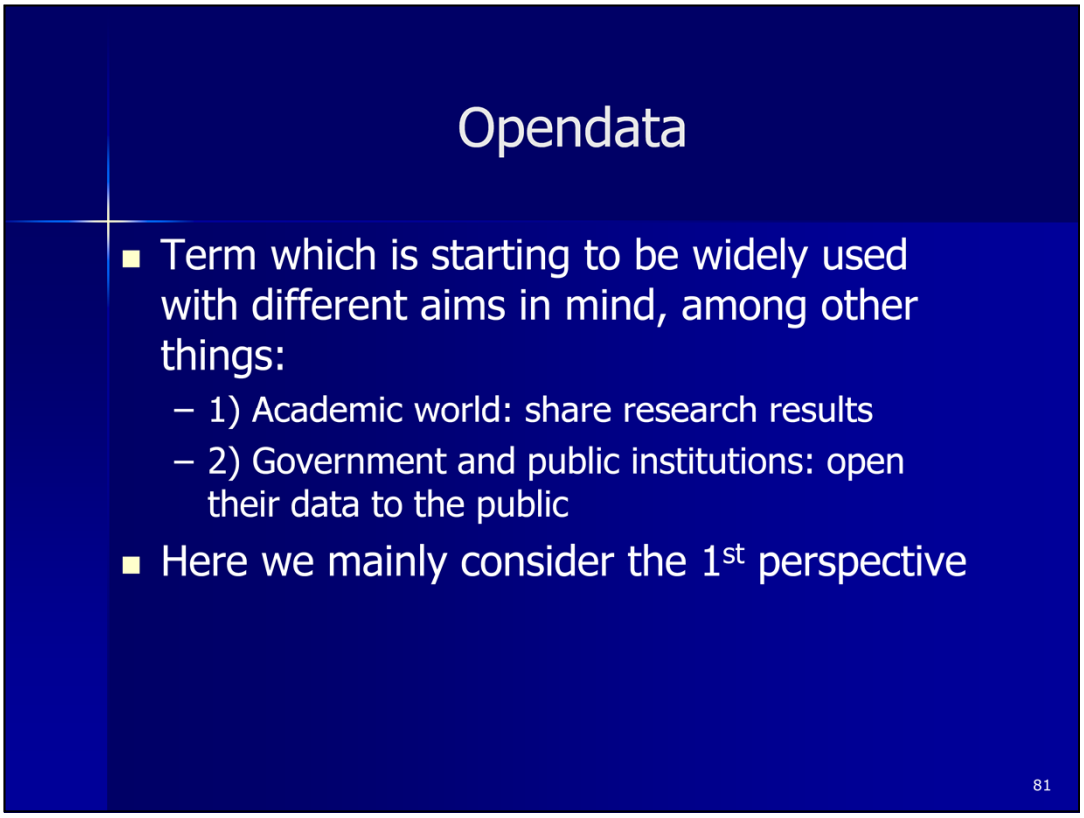
Here is a recent statement from the European commission which give guidelines to researchers for the "Horizon 2020" framework (recently opened).

The report says : there should be open access to publications resulting from publicly founded research as soon as possible, and forms should be changed in order to let researchers retain their copyright while granting licences to publishers. These are 2 different issues. Open access needs not wait for the new wording of copyright forms.

There exist a huge literature (open access) on this topic, which details the main ways of providing open access to publications, the so-called Green and Gold roads. As regards CALL journals and the Gold Road, Language Learning & Technology (LLT) and Alsic were the first to provide open access to their articles. Recently CALICO reduced its moving wall to one year (one year after their publication, articles become open access).

Let me just recall here, it is firstly the author responsibility to deposit her/his article once it has been accepted by reviewers in open archives (the green road), whichever kind of archive it is, national or institutional.

Let us spend a little more time on access to research data, to the new concept of OpenData



This term starts being widely used with different aims in mind:

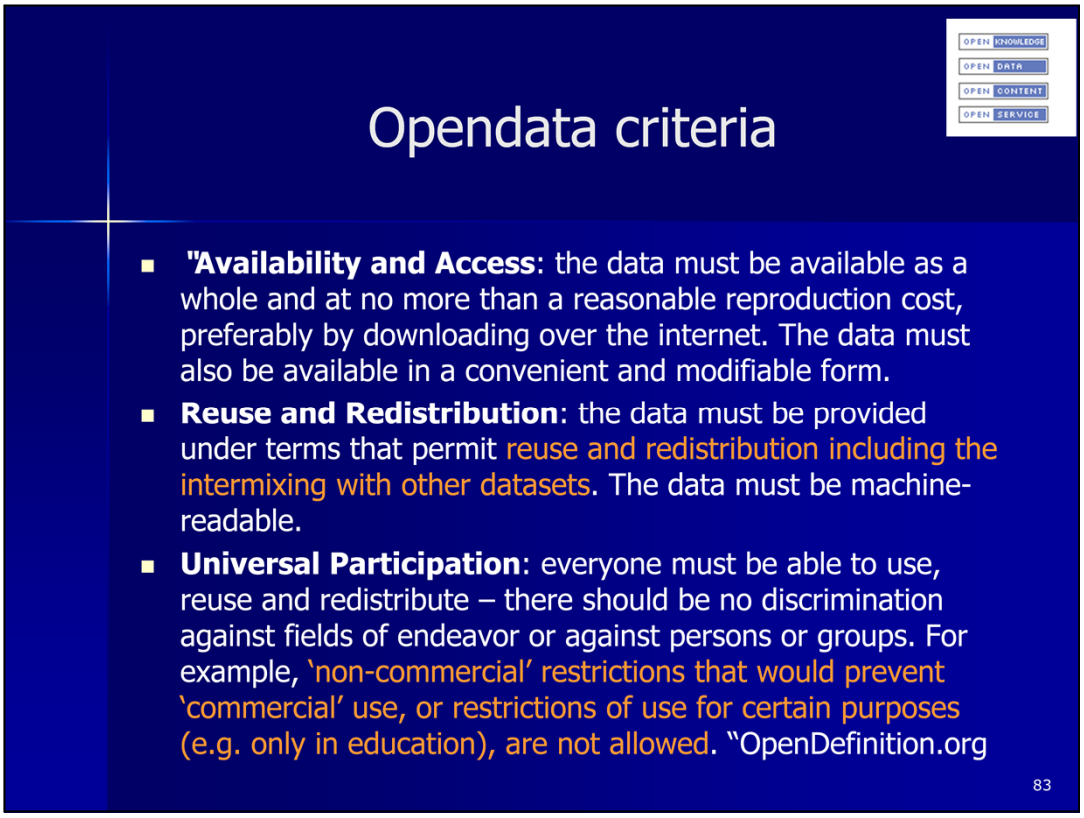
- In the academic world it refers to the way of sharing research data
- For government and public institutions, they started opening their data to the public.

This is of course the first case we will here consider.

Opendata def

- "Open data is data that can be freely used, reused and redistributed by anyone – subject only, at most, to the requirement to attribute and sharealike." OpenDefinition.org

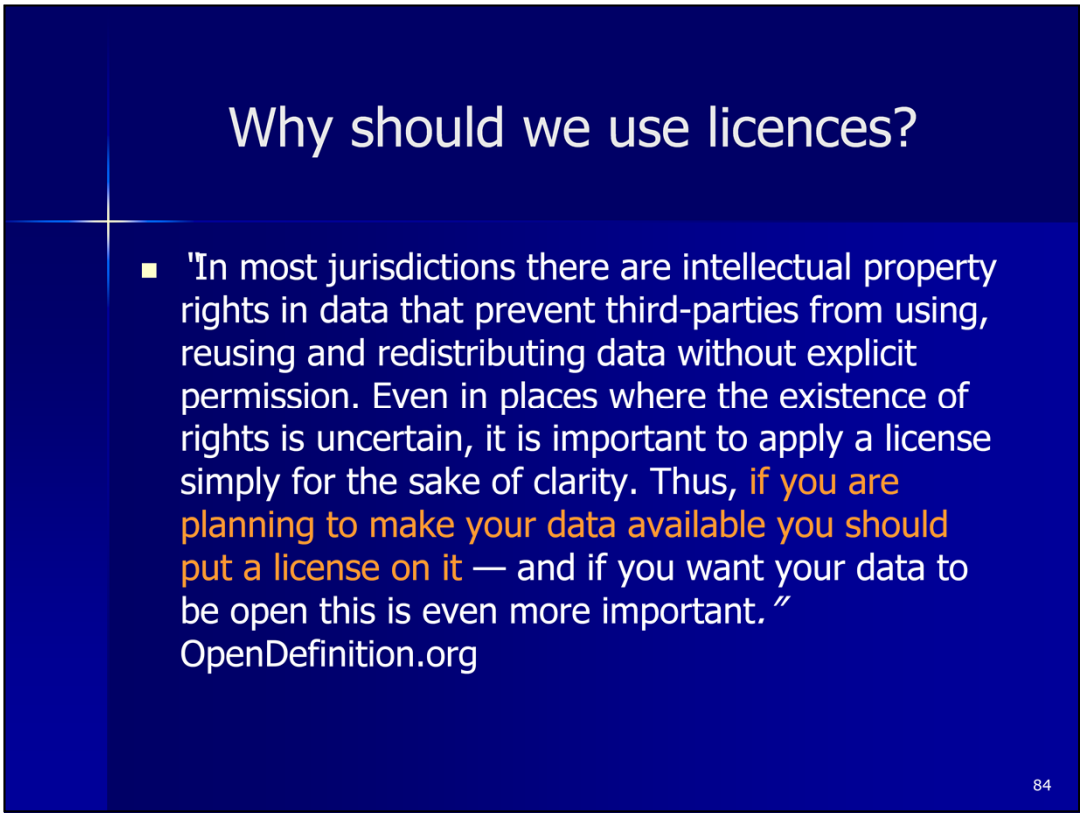
82



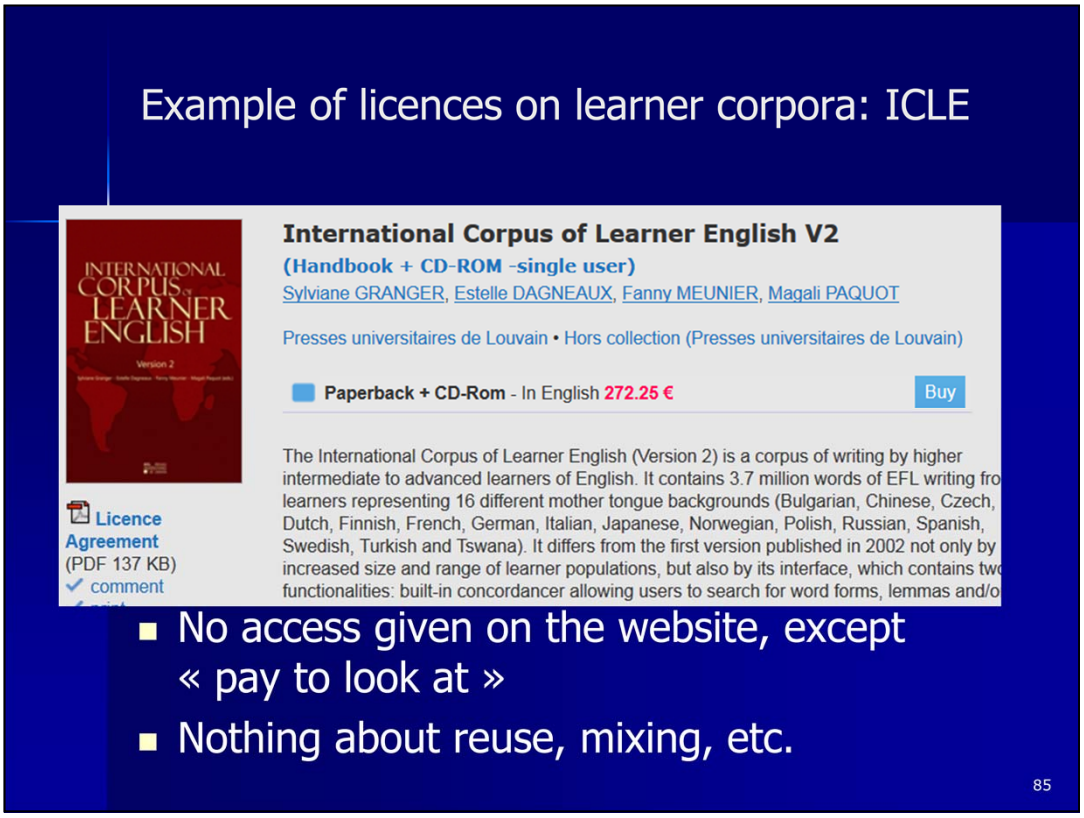
There exist 3 main criteria that research data should follow in order to be considered OpenData.

Besides being obviously available, the interesting perspective is the fact that data can be access in order to be reuse and mix with other data.

Second interesting point is that the constraints for reuse should be reduced to a minimum, then the definition stipulate that non-commercial' restrictions that would prevent 'commercial' use, or restrictions of use for certain purposes are not allowed



Another interesting point is also the fact that authors should always put a licence on data, when they plan to make their data available.



What about access to data coming from the language learning fields. Here is the example of the most well known learner corpus, namely the International Corpus of Learner English.

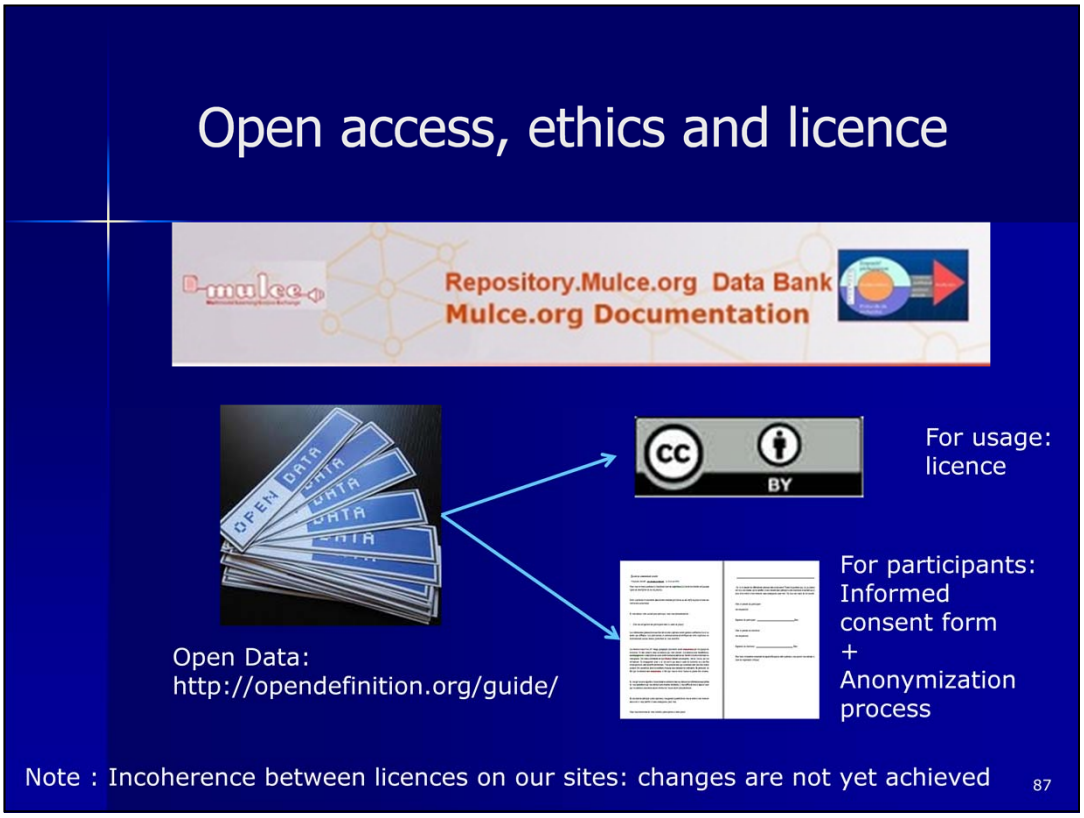
On the website, no access is given, except the “pay to look at”. Nothing is mentioned about reuse and mixing.

Here is another learner corpus , the ELFA (English as a Lingua Franca in Academic Settings).

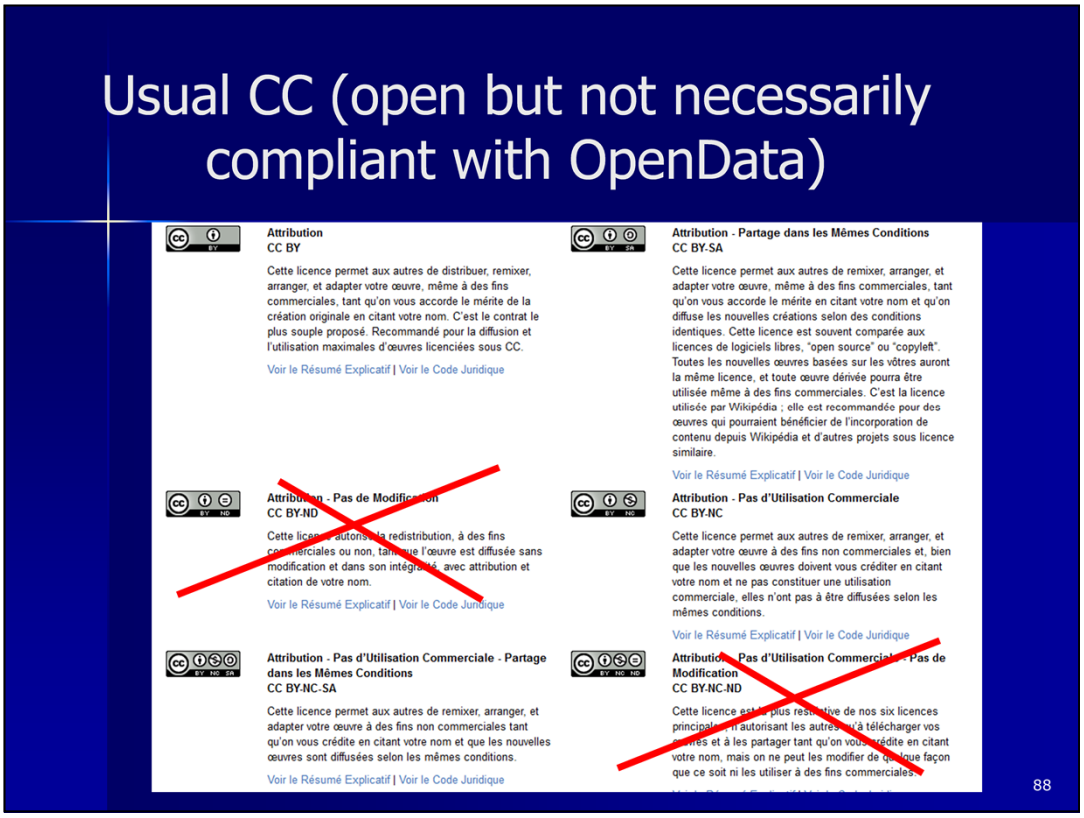
Creators made some progress.

In the earlier license, users were being charged over 100 euros for a mere six-month license (and just for the text corpus, not the audio), with instructions to "destroy" the files at the end of this period or purchase a new license!

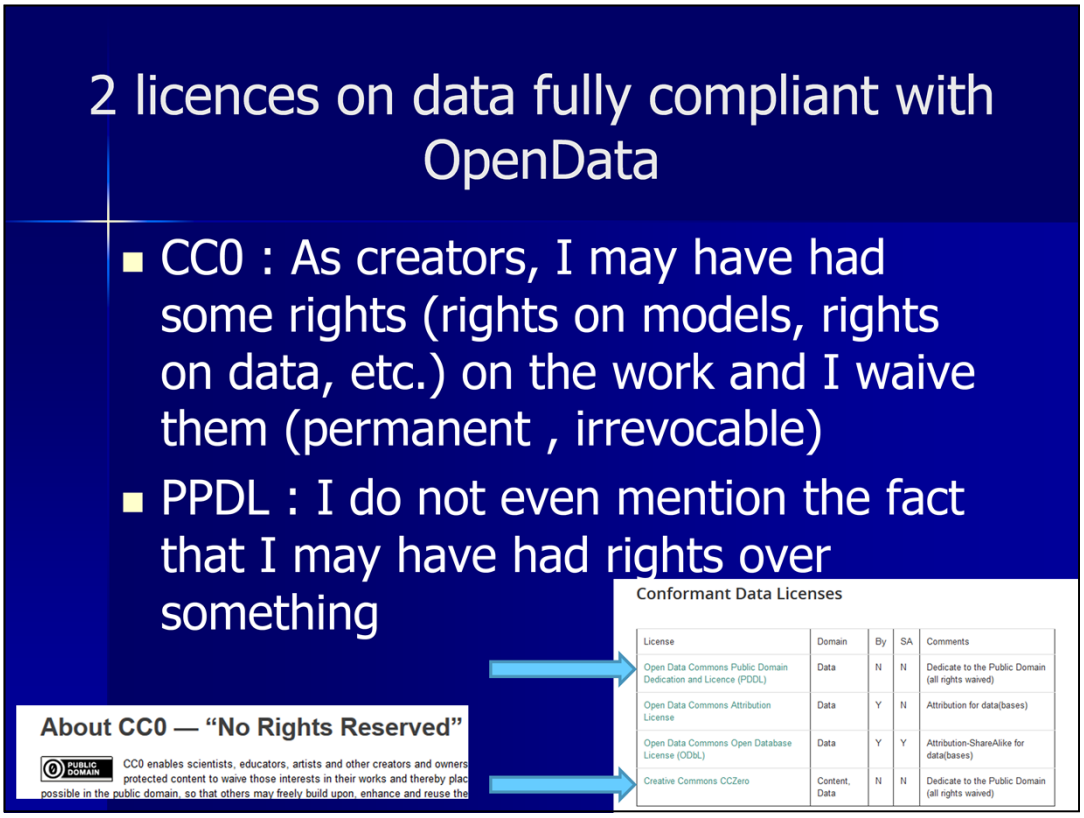
Now text data are open access but for personal use only (i.e. not here for example). They are important restrictions in the license and still no access is given to the audio.



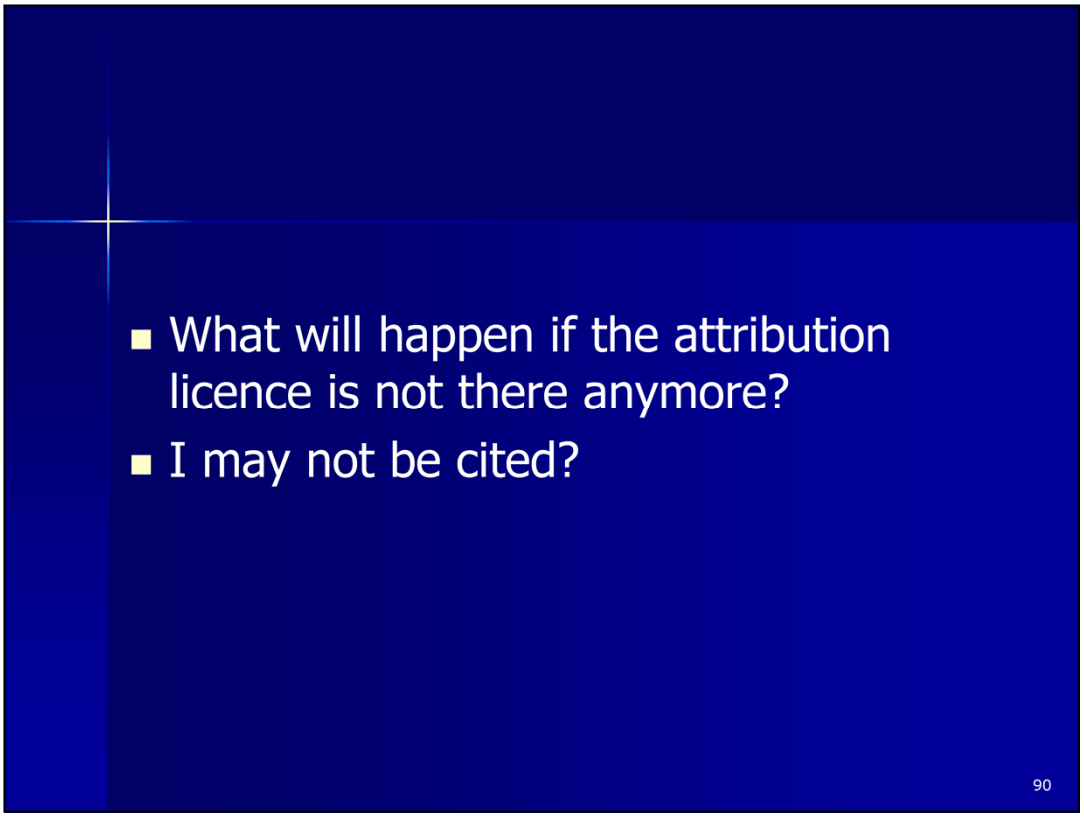
As regards the LETEC corpora and Mulce repository, all our corpora are of course open access , without any registration. In each corpus we included the informed consent form signed by participants. We started to change our licence in order to become fully compliant with OpenData criteria.



Here is the listing of the Creative Common licences, created by James Boyle and his colleagues. 2 of them are not OpenData compliant, because they forbid direct use for commercial purposes. In Mulce, we started to switch from the BY-NC-SA to the BY only, i.e. the only obligation of to refer to the authors and the original work.

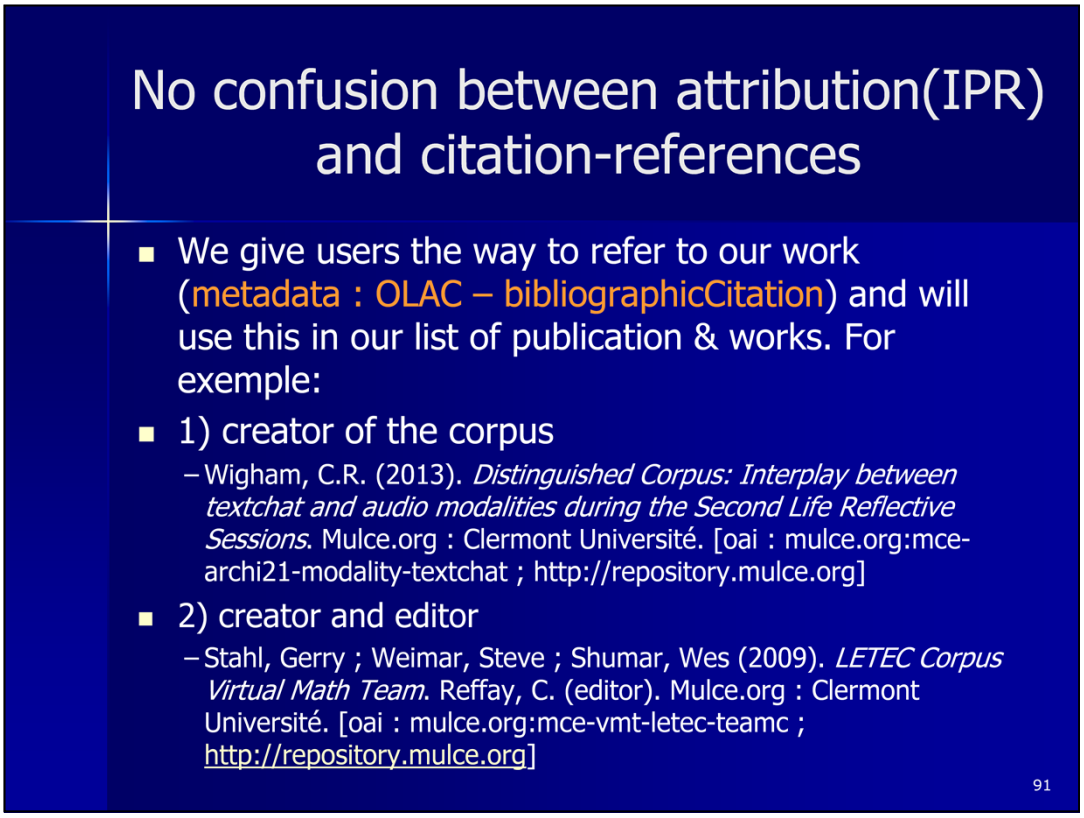


The CC-BY licence is an improvement. But it still restricts possibilities for mixing and reusing data. There exist 2 licences which are fully compliant: the PDDL and CCO ones. These licences make a significant step forward. They wave intellectual property rights, data then become part of public domain.



When hearing this you may be afraid or at least sceptical:

- What will happen if the attribution licence is not there anymore?
- I may not be cited?



But we should not be afraid.

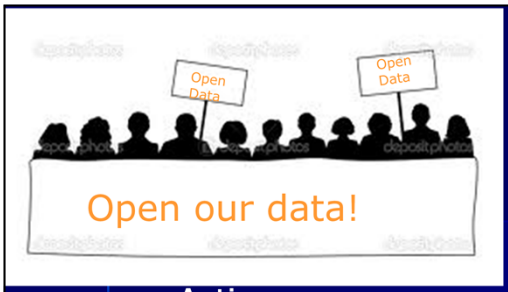
We have the habits of confusing 2 very different issues IPS and citation-references:

- The first one only refer to legal issues : “you did not cite me, I am going to take you before the court!”
- In the second one, we have our academic procedure. We need to refer to previous work and when authors do not do it properly, their work is rejected by peers-reviewers.

Hence we need only worry about correctly referencing our work and making this reference clearly available.

Here are for example two references toLETEC corpora, the first one to the author, only creator of the corpus ; the second one where creators and editors are distinguished (like a chapter of a book).

Moreover these references are tagged as such, included in metadata which are harvested on Internet thanks to the OLAC harvesting protocol.



Recommendations

If we want to be connected to Digital Humanities

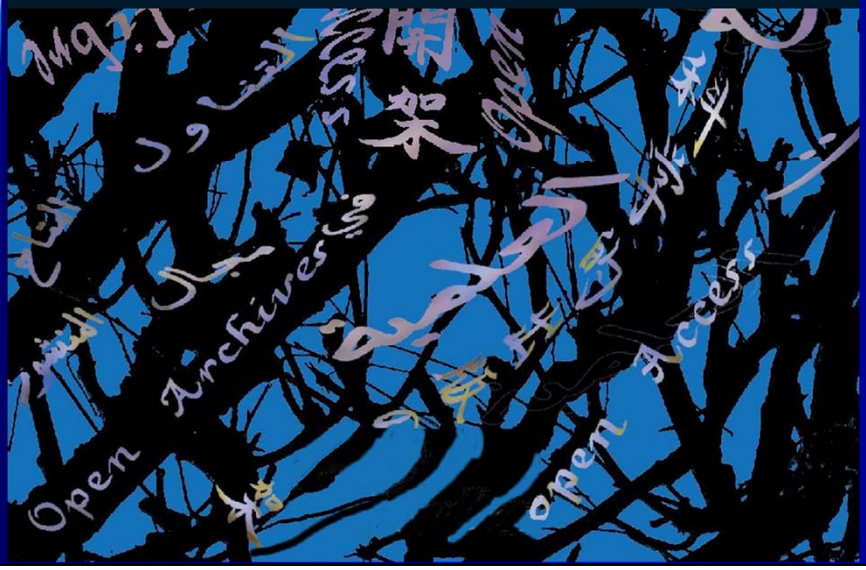
- Actions
  - Open our data (provided that ethics is OK – anonymisation)
  - Choose licences with the fewest restrictions
  - Cite others and your data as bibliographic references
  - List them in your work
- Implications
  - Acknowledgement will come (from institutions, other colleagues)
  - CALL research will progress (re-analysis, coverage extended with mixing)
  - CALL data will be reused by other fields

92

Thank you for your attention!

Thierry.chanier at univ-bpclermont.fr

http://lrl.univ-bpclermont.fr/spip.php?rubrique98



93