Istituto di Linguistica Computazionale
(ILC) CNR of Pisa

Industrial Philology: problems and techniques of data and archives preservation for future generations

E. Sassolini, A. Cinini, S. Sbrulli, N. Cucurullo, and M. Sassi

Industrial Philology

ILC fifty-year history of ICT applied to the NLP

wide variety of texts and corpora that have been stored in various formats and record layouts

Cultural Heritage domain

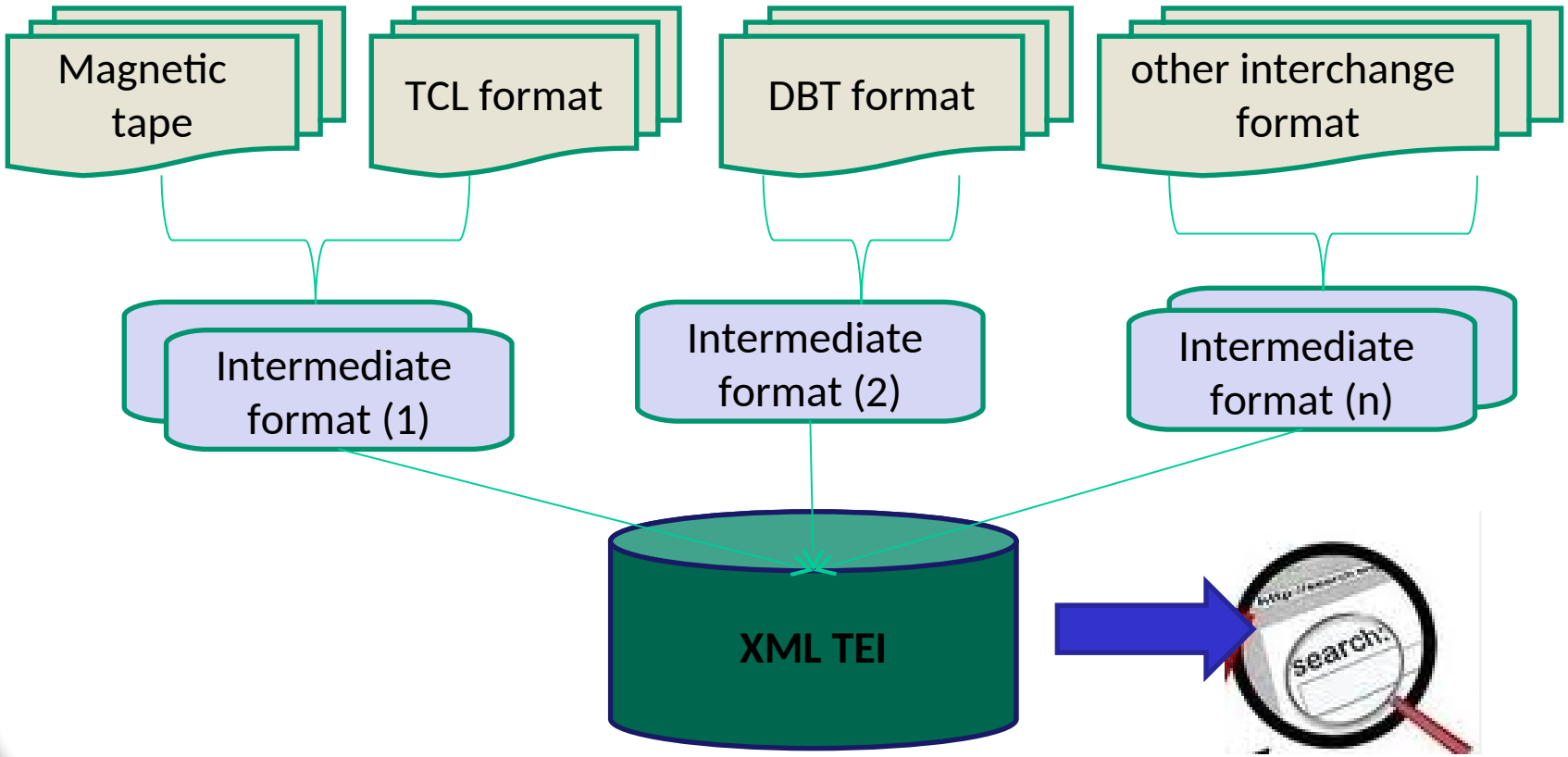
e.g.: texts in Latin and ancient Greek, which required a complicated system of encoding in the 60s-70s punch cards, with a limited set of characters available were used

equivalence tables were created, so that a full performance can be obtained using Unicode encoding.

text handling problems

- character encodings, based on different operating systems used over the years (from EBCDIC, passing by ASCII, to Unicode)
- textual materials produced in or recovered from a project of the past do not have a standard format, but are the expression both of technology and of the research needs at that time

Digital Text Repository schema



Text acquisition strategy

Source text	Transition phases (TP) required	Meta data
Text in magnetic tape	Many type TP	study and research in the archives ILC
Text divided into separate resources	TP>3	recovered from paper-based data
Text in file obsolete	TP>2	recovered from paper-based data
Text digital with obsolete character encoding	2<TP<3	recovered from: - paper-based data - the digital format
Digital text	One TP	recovered from the digital format

Annotated text acquisition strategy:

Source text	Transition phases (TP) required	specific annotations type encoding	Meta data
Text in magnetic tape	Many type TP	?	work long and difficult
Text divided into separate resources	TP>3	DBT type encoding	recovered from paper-based data
Text in file obsolete	TP>2	Obsolete type encoding	recovered from paper-based data
Text digital with obsolete character encoding	2<TP<3	Specific type encoding	recovered from: - paper-based data - the digital format
Digital text	One TP	ILC text encoding	recovered from the digital format

Phase 1: texts material collection

- Research of all existing text materials in ILC, looking to the historical projects to which ILC has worked;
- Identification of physical locations where these texts are;
- Recovery and analysis of text material;
- Quantification of the work required.

Phase 2: text corpus standardization

- For each type of textual data definition of:
  - ✓ a procedure for the text standardization;
  - ✓ Xml TEI model of text representation
  - ✓ costs of such work

Tools

- Tools of texts analysis:
  - ✓DBT (Data Base Testuale);
- Modules and procedures for specific text recovery;
- Converter and parser XML TEI for text corpus testing

projects & applications

Scientific Cooperation Agreement between the ILC (Computational Lingusitics Institute) and the "Accademia della Crusca" of Florence

- ✓ Selection of relevant text materials in ILC and then specific classification;
- ✓ Identification of the text encoding and, where present, of the linguistic annotations associated;
- ✓ Conversion of texts into a shared and standardized representation format;
- ✓ Development of a text management system for the advanced search functionalities.

Open questions

To preserve the ancient procedures in programming languages which printouts of processing still exist:

- Can they be preserved or should they be dropped?
- Can they be considered a form of “industrial Philology” and maintened?