# Timelink: Conceitos básicos para gestão do processo de transcrição

## As fases do processo de tratamento de fontes histórias no Timelink

4 fases no processo de tratamento dos dados: *transcrição, tradução, importação e identificação*.

*Transcrição* corresponde à codificação da informação das fontes numa notação formal que permite o seu processamento pelo computador.

A *tradução* consiste no processamento da transcrição para extração de dados para importação na base de dados.

A *importação* é a operação pela qual o resultado da tradução é inserido na base de dados.

A *identificação* é a operação pela qual se agregam diferentes ocorrências de pessoas nas fontes em biografias. O sistema distingue uma "ocorrência", que é uma referência a uma pessoa num documento ou obra, de uma "pessoa real", que é a pessoa que efetivamente existiu e deixou traços nas fontes. Identificar é juntar as ocorrências de cada pessoal real.

A identificação é um processo em aberto, em contínua revisão, e é feita de duas maneiras: anotando as ocorrências na fonte com o elemento "mesmo_que", que indica que uma ocorrência corresponde à mesma pessoa que outra e "ligando" as pessoas no interface da base de dados que fornece várias oportunidades agregar ocorrências em pessoas reais.

Em qualquer momento as identificações podem ser exportada sob a forma de ficheiros que podem ser trocados entre instalações e pessoas.

## Diferentes tipos de ficheiros

O processo descrito acima produz vários tipos de ficheiros de texto.

- Ficheiros em formato Kleio, com extensão .cli. Contêm as transcrições de fontes e e informação auxiliar, como listas de pessoas identificadas.
- Ficheiros de dados para importação, com formato .xml, contêm o resultado das tradução dos ficheiros .cli num formato apropriado para importação na base de dados.
- Relatórios de tradução .rpt e importação .xrpt são ficheiros de texto com relatórios sobre os processos de tradução e importação e eventuais mensagens de erro ou aviso.
- Ficheiros auxiliares de gestão do programa (.str,.srpt, .err, .xerr, org, old, ids).
- Cópias de segurança da base de dados (.sql)

O utilizador trabalha essencialmente com ficheiros kleio, de extensão .cli, usando um processador de texto para textos codificados (atualmente o Visual Studio Code da Microsoft, com a extensão Time Link Bundle).

Todos os outros ficheiros são gerados pelo programa durante as diferentes fases de processamento.

## A cada projeto tem um repositório Git de referência.

A cada projeto Timelink corresponde um repositório Git de referência, que contém todos os ficheiros gerados pelo processo de transcrição, tradução, importação e identificação.

Git é uma ferramenta criada para gerir o conteúdo de ficheiros que se alteram ao longo do tempo pelo contributo sucessivo de várias pessoas.

> O Git é tipicamente usado em projectos informáticos em que o resultado final é um programa de computador. Nesses projectos é necessário integrar contributos de várias pessoas e alterações frequentes a ficheiros, devido a correção de erros ou implementação de novas funcionalidades.

> No Timelink o Git é utilizado para controlar os ficheiros com transcrições de fontes históricas, recolhidos por uma ou várias pessoas, e as identificações de co-ocorrências de pessoas e entidades nas fontes. O resultado final é uma representação da informação contida nas fontes históricas sob a forma de base de dados pesquisável.

Quando se inicia um novo projecto, ou se migra um projeto existente para este formato, é criado um repositório Git acessível via internet.

### Em cada repositório o ramo master corresponde à versão de referência.

Os repositórios Git podem conter vários *ramos* que correspondem a versões criadas ou alteradas por diferentes intervenients no projeto.

O ramo master preserva a *versão de referência* de todos os ficheiros produzidos pelas várias fases do processo. O ramo master alimenta uma base de dados de referência para a comunidade, e pode alimentar outras cópias que sejam criadas para comodidade e redundância.

> Isso significa que é possível duplicar todo o projecto fazendo um clone unidirecional do ramo master e refazendo a importação das fontes e das identificações para uma base de dados alternativa.

A gestão do ramo master cabe ao responsável do projecto.

### Se há mais que uma base de dados, cada deve estar associada a um ramo.

Como ficou acima, a base de dados de referência de um projeto está associada ao ramo master do repositório de projeto.

Também é possível fazer uma variante da versão de referência, acrescentando mais fontes, alterando formatos de transcrição e regras de inferência e identificando pessoas e entidades de forma diferente.

Nesse caso cria-se um novo ramo a partir do principal, ou faz-se uma bifurcação. e altera-se como for necessário.

### Transcritores trabalham sobre repositórios isolados

..... TBD-----

# Conceitos base e enquadramento geral

## Comunidades, fontes e modelos de análise

No Timelink tratam-se conjuntos de fontes que correspondem a *comunidades*, em sentido lato.

Nos casos mais típicos as comunidades são uma localidade, uma paróquia, uma cidade, uma empresa.

No sentido mais geral a comunidade é um conjunto de pessoas e outras entidades que têm intereações frequentes entre si e por isso deixam nas *fontes históricas* uma série de registos que podem ser correlacionados e estudados, reconstruindo as *histórias de vida* e as *redes de relações* das pessoas e entidades envolvidas.

Um projeto elaborado com Timelink envolve a transcrição de fontes históricas em quantidades significativas, seguido do seu processamento para alimentar uma base de dados centrada nas pessoas, os seus atributos, as suas relações e as funções que assumem em diferentes fontes.

## Transcrição, tradução, importação e identificação

Cada comunidade tratada com a metodologia Timelink produz uma série de representações informáticas que correspondem a níveis progressivos de abstração, partindo da informação das fontes até à elaborações de modelos analíticos pelo investigador, como *histórias de vida*, *redes interpessoais* e outras construções intepretativas.

Essas representações informáticas são as seguintes:

- um conjunto de *transcrições* de fontes históricas numa linguagem formal chamada kleio.
- um conjunto de *traduções*, que são ficheiros produzidos a partir das transcrições e que preparam a informação para importação numa base de dados.
- um conjunto de *interpretações*, em especial decisões tomadas pelo investigador sobre a co-ocorrência de pessoas, bens e outro tipo de entidades. Ao decidir que diferentes referências em diferentes fontes dizem respeito a uma mesma *entidade*, o investigador permite a geração de representações derivadas como biografias, redes, etc... que suportam análises complexas da comunidade. Identificar as mesmas pessoas em várias fontes diferentes não é uma operação determinística. Diferentes investigadores podem tomar decisões diferentes com base na sua análise dos dados e a importância que dão a semelhanças e diferenças específicas na informação disponível. Por isso considera-se a *identificação* de pessoas uma interpretação.

## O processo e os seus produtos

Estas representações informáticas resultam de um *processo* que vai desde a atividade determinística de *transcrição* até às decisões complexas, e por vezes ambíguas, da *identificação*. Em cada fase do processo são feitas transformações específicas da informação, utilizando ferramentas que produzem ficheiros com informação progressivamente mais *complexa* e *menos determinística* (no sentido em que é progressivamente mais afetada por escolhas e decisões a partir da informação factual). O processo permite definir diferentes *papéis* e *responsabilidades*, que num projeto pequeno pode ser assumidos sempre pela mesma pessoa, mas que num processo maior podem envolver várias pessoas articuladas.

1. O *transcritor* regista o conteúdo das fontes segundo um formato pré-determinado que procura captar a informação numa forma próxima do texto original.
    - As principais decisões que toma são de natureza paleográfica e de compreensão do formato e notação de transcrição.
    - Em geral os formatos de registo pré-definidos não permitem uma grande variabilidade de resultados. Diferentes *transcritores* perante a mesma fonte produzem textos idênticos, se não fizerem erros de leitura.
    - O resultado da transcrição são ficheiros com a extensão cli. O conjunto dos ficheiros cli de um projecto constituti a *base factual* (as fontes primárias) de um do estudo de uma comunidade.

2. O *tradutor* gere o processo de tradução das fontes transcritas.

    - Esta fase é feita por um programa informático que incorpora uma série de decisões prévias sobre a forma de representar a fonte (a sua *estrutura* ou *formato*) e também um conjunto de *regras de inferência* de informação implícita na fonte.
    - As regras de inferência permitem aliviar o trabalho de transcrição ao inferirem atributos como o *género*, *estado civil*, assim como *relações de parentesco* e outras, a partir das funções com que pessoas e outras entidades ocorrem nos actos.
    - Cabe ao *tradutor* aferir se o formato usado para a transcrição da fonte encapsula o máximo de informação relevante sem custo exagerado de transcrição.
    - Cabe também ao *tradutor* determinar as regras de inferência aplicáveis a cada tipo de fonte, seguindo o princípio minimalista de não inferir o que pode ser ambíguo ou sujeito a discussão. As regras de inferência visam sobretudo evitar transcrição de informação redudante pelos *transcritores*.
        - Por exemplo, num batismo não vale a pena registar o sexo da mãe e do pai, nem a relação de parentesco entre pais e criança batizada - tudo isso pode ser inferido automaticamente.

    - Finalmente o processo de tradução gera identificadores únicos para cada entidade (pessoa, bem, etc.) referida na fonte, identificadores esses que são necessários para a posterior identificação das ocorrências das mesmas entidades nas fontes.
        - A geração de identificadores é um processo automático mas o *tradutor* pode definir alguns parâmetros que regulam o processo: quais os elementos da fonte que terão identificadores gerados automaticamente e quais os que requerem a atribuição de um identificadores pelo *transcritor* e ainda se os ids num dados ficheiro devem ser prefixados automaticamente com uma sequência de caracteres.

    - O resultado do processo de tradução são diferentes ficheiros, com o mesmo nome que o ficheiro com a transcrião da fonte e com diferentes extensões:
        - xmlcontêm os dados retirados do ficheiro cli num formato adequado para importação na base de dados.
        - rpt contêm relatórios de tradução para deteção de erros ou situações que necessitam cuidado(*warnings*).
        - err contém o número de erros e avisos gerados pela tradução
        - cli o ficheiro cli original é regravado com uma formatação estruturada para melhor legibilidade e com os identificadores únicos de cada entidade referida incluídos sempre que necessário para que futuras re-traduções e importações não invalidem identificações feitas com base nos identificadores gerados na tradução.

    - Adicionalmente, o resultado de uma tradução é determinado por dois tipos de ficheiros que controlam o comportamento do tradutor e que tipicamente são os mesmos em todas as transcrições de um projeto.
        - O ficheiro de extensão str (normalmente gacto2.str) que define o formato de transcrição da fonte (que tipo de actos, qual a informação associada a cada um e que tipo de actores participam, com que funções).
        - O ficheiro inferences.plcom as regras de inferência da informação implícita na transcrição.

- Assim, o resultado final da tradução das fontes é dado pelo conjunto dos ficheiros .cli, .rpt, .err, .xml e pelos ficheiros gacto2.str e inferences.pl.

3. O importador incorpora o resultado da tradução numa base de dados que vai permitir a identificação das pessoas.

    - A base de dados permite navegar de forma eficaz os dados contidos nos ficheiros xml gerados pelo *tradutor*.
    - É importante entender que no Timelinka informação importada é imutável. Não existe modo através do interface da base de dados de alterar os dados importados.
    - Quando um erro é detectado na base de dados a correção tem de ser feita na transcrição original, que é re-traduzida e reimportada. Este princípio garante a transparência da informação usada no projeto e a sua acessibilidade na forma de transcrição de fonte.
    - Assim a fase de importação não adiciona informação à fase anterior. Apenas a transforma num formato mais acessível.
    - Esta fase pode ser representada por uma exportação em linguagem sql do conteúdo da base de dados e pelos ficheiros .xpt com o relatório de importação e .xerrcom o número de erros detectados no processo de importação.

4. O identificador (ou o *investigador*, porque pode gerar também redes e grupos) toma decisões sobre quem é quem na informação recolhida e pode gerar entidades derivadas como redes e grupos.

    - Na sua essência o processo de identificação regista decisões do tipo: a pessoa X que ocorre no acto A é a mesma que a pessoa Y que ocorre no acto B.
    - Essas decisões são registadas em tabelas específicas na base de dados e podem ser exportadas em formato kleio , facilitando a troca de dados.
    - Como as identificações são feitas na base de dados elas também são incluídas em ficheiros de exportação da base de dados em formato sql.