Migração de Dados do Sistema Científico Português para a Plataforma Lattes

L. Santos, L. Amaral e J. Oliveira
Departamento de Sistemas de Informação
Escola de Engenharia, Universidade do Minho
4800-058 Guimarães, PORTUGAL
leonel@dsi.uminho.pt, amaral@dsi.uminho.pt, jno@dsi.uminho.pt

Resumo

Este artigo tem como finalidade descrever o processo de migração de dados do Sistema Científico Português para a Plataforma Lattes. Esses dados estavam distribuídos por dois organismos do Ministério da Ciência e Tecnologia, o Observatório das Ciências e Tecnologias (OCT) e a Fundação para a Ciência e Tecnologia (FCT). Este processo revestiu-se de alguma complexidade devido às múltiplas Bases de Dados existentes sobre as mesmas entidades do sistema científico nacional e com informação também ela diferente, uma vez que se destinavam a diferentes finalidades e cuja recolha tinha sido realizada em momentos diferentes. Por estas razões o mapeamento da informação numa Base de Dados única era uma tarefa de grande dificuldade atendendo ao reduzido tempo disponível para a sua concretização.

Após a decisão de adesão de Portugal à Plataforma Lattes e à Rede SCienTI, ficou claro que o sistema teria que ser lançado com os curricula vitae mínimo de todos os investigadores portugueses registados no Ministério da Ciência e Tecnologia, razão pela qual o sucesso deste sistema era vital para a participação portuguesa na rede.

Palavras Chave:

Ciência e Tecnologia, Sistema Científico Português, Bases de Dados, XML, Serviços de Informação on line

Introdução

O sistema científico português, em Fevereiro de 2002 era tutelado pelo Ministério da Ciência e Tecnologia (MCT) e que pode ser visto com mais detalhe em [Amaral et al. 2002]. No que ao processo de migração de dados interessa, existiam dois organismos que dispunham, e forneceram, informação para que fosse incluída na versão inicial da Plataforma Lattes, versão portuguesa, a finalidade era lançar a plataforma com um conjunto de *curricula vitae* representativo da comunidade científica nacional, esses organismos foram a Fundação para a Ciência e Tecnologia (FCT) e o Observatório das Ciências e Tecnologias (OCT).

Este artigo visa descrever a fase inicial do processo de adopção da Plataforma Lattes em Portugal – A migração de Dados.

O processo foi realizado na segunda metade do mês de Fevereiro do ano 2002, por uma equipa da Universidade do Minho com a colaboração da universidade de Trás-os-Montes e Alto Douro, sendo constituída por quatro técnicos formados em Engenharia informática.

A equipa sempre esteve consciente das pressões temporais para concluir a migração de dados e também sempre teve consciência da má qualidade das bases de dados existentes. Contudo a qualidade dos dados foi sempre uma preocupação que balizou todo o processo.

O Problema

O problema que foi colocado à equipa foi o de migrar os dados existentes em diversas bases de dados do Ministério da Ciência e Tecnologia, nomeadamente no OCT e na FCT para a Plataforma Lattes. Essas bases de dados continham informação sobre os Investigadores, as Unidades de Investigação, os Projectos de Investigação financiados pela FCT e as Publicações Científicas referenciadas no ISI e complementadas com publicações das áreas

científicas das ciências sociais e humanas. A informação relativa a cada uma destas entidades estava presente nos dois organismos e em diversas bases de dados. Muita da informação existente tinha sido recolhida em diferentes momentos temporais e para diferentes fins, o que implicava que a informação recolhida era muitas vezes diferente para as mesmas entidades. O que se traduzia em redundância de informação, informação incompleta e nalguns casos a sua inconsistência.

Além destes problemas havia que conhecer a arquitectura dos dados do Lattes e compatibilizar os dados do sistema científico português com essa arquitectura.

Outras dificuldades que se colocaram à equipa foram a terminologia utilizada nos dois países que nem sempre se resumia a uma tradução directa e as diferenças existentes na organização dos sistemas científicos de Portugal e do Brasil.

Este foi o problema colocado à equipa e que se esperava que fosse terminado com sucesso em duas semanas, do qual dependia a adesão de Portugal à Plataforma Lattes e à Rede SCienTI.

Porquê a migração dos dados para a Plataforma Lattes e não a sua adopção e lançamento sem informação, sendo a mesma introduzida pelos próprios investigadores à medida que iam aderindo à plataforma?

A razão prende-se com a nossa experiência anterior no SICT – Sistema de Informação de Ciência e Tecnologia. Este sistema foi concebido e desenvolvido em Portugal no âmbito do Projecto GEIRA de 1996 a 1999 por um grupo de investigadores do Departamento de Sistemas de Informação da Universidade do Minho e da Universidade de Trás-os-Montes e Alto Douro. O seu objectivo era a criação de um sistema de gestão de curricula dos investigadores portugueses.

Essa experiência demonstrou que para que o sistema atinja rapidamente massa critica que o leve ao sucesso, é necessário,

entre outros factores, que as pessoas acreditem que aquele vai ser "O Sistema" de suporte ao Sistema Científico e que não é mais um sistema experimental. Por outro lado não faz sentido que um organismo solicite aos seus agentes informação de que já dispõem, mesmo que tenha sido recolhida para outros fins. É o princípio do combate à burocracia.

A qualidade da Informação

A gestão dos dados é das áreas que menor atenção recebe no contexto organizacional, e é também aquela que tem menor sustentação de um saber teórico-prático. A discussão pode mesmo iniciar-se pelo apuramento do significado da trilogia dados-informação-conhecimento, exercício esse que muitas vezes extravasa a utilidade dos seus resultados. Nesta abordagem vai-se ignorar a distinção entre dados e informação, até porque se informação é um conjunto de dados organizados de acordo com uma dada racionalidade, racionalidade esta definida por um consumidor, essa mesma informação passa a dados quando esta é por sua vez *input* de uma outra racionalidade. É necessário reparar que quer seja apelidada de dados ou de informação, está-se a falar dos mesmos signos, por vezes nos mesmos suportes.

O reconhecimento da informação como algo que pode ser gerido é dificultado por várias razões. A informação não é um recurso como outro qualquer. Não é fácil inventariar a informação que se tem, ou conhecer os custos de gestão por não ter informação ou por esta ser deficiente. Conceitos como custo, valor ou qualidade, não são aplicados à informação como são relativamente a máquinas ou a produtos. Um outro aspecto é a confusão entre informação e tecnologia. Para muitos, informação é algo tecnológico, não sendo encarada como uma responsabilidade de gestão mas como algo a entregar aos informáticos.

Apenas uma razão bastaria para demonstrar o quanto errada está esta visão, e essa seria os custos em que se incorre quando se "faz mal" por não ter informação adequada para trabalhar:

- Processos de decisão afectados
- Dificuldade em extrair informação para actividades de gestão
- O custo de resolução de erros
- Projectos como *datawarehousing* ou migração de dados afectados
- Perda de confiança
- Desmotivação

O facto de não ter informação com qualidade faz com que os processos que a utilizam não possam ter um desempenho adequado, ou limita a própria natureza da arquitectura de processos que está na base de uma actuação ou política.

Importa agora questionar o que é a qualidade da informação e qual é a tradução operacional desse conceito. [Juran e Gryna 1993] definem qualidade como "*fitness for use*". Esta definição possui a vantagem de ser simples e explícita num aspecto extremamente importante da qualidade, que é o facto da qualidade não existir apenas por si só nas características intrínsecas dos "objectos", mas também na utilização ou aplicação desses objectos. Assim, como a qualidade de um televisor ou de uma ferramenta de *software* só pode ser sentida por quem o usa, a qualidade da informação só pode ser avaliada por quem a consome, e só terá qualidade se quem a consome a considerar como apropriada para as necessidades em causa. E se facilmente encontramos características num televisor, como a facilidade de configuração ou as dimensões físicas, também na informação poderemos encontrar características determinantes para o seu "*fitness for use*".

As características da qualidade da informação foram trabalhadas por vários autores [Wang et al. 1994; Redman 1996, Barquin e Edelstein 1997]. Desse trabalho, emergem várias considerações. Em primeiro lugar, a qualidade da informação vai muito mais além do que as questões da correcção, ao surgirem características como a actualidade, a utilidade ou a compreensão. Em segundo

lugar, a separação entre as características intrínsecas da informação e as derivadas do seu consumo. Por exemplo, a informação pode estar correcta mas ser irrelevante. Por último, a própria qualidade do modelo conceptual pode ser objecto de avaliação. Segundo [Watson 1999], dois critérios existem para julgar a qualidade de um modelo de dados, se o modelo está "bem feito", ou seja, se obedece às regras da modelação e se não há ambiguidades, e se a imagem da realidade representada traduz com detalhe e correcção a realidade que foi modelada. Moody [Moody 1998; Moody e Shanks 1998] vai muito mais longe, ao definir um conjunto de dimensões para a qualidade de um modelo e métricas para essas dimensões, e ao determinar várias actividades de análise e de inspecção do modelo de dados e a respectiva localização no processo de desenvolvimento de sistemas de informação.

As habituais limpezas de dados são abordagens de inspecção, pois procuram identificar os defeitos nos dados depois de estes terem sido criados. O termo deficiência de dados é explicado em [Wand e Wang 1996] como a discrepância entre a visão que o utilizador tem do mundo real e a visão dessa mesma realidade que é transmitida pelo sistema de informação. Para de alguma forma controlar este problema, algumas técnicas, mesmo automáticas, auxiliam a detecção e mesmo a correcção da informação. Esta solução é de curto prazo, pois os processos que originaram o erro mantêm-se inalterados. É necessário inverter a tendência de inspecção pela da prevenção, ou seja, impedir a reincidência do erro, caminho este preconizado pelas abordagens à qualidade. Esta posição exige a procura da origem dos erros, a sua eliminação e posterior monitorização.

A *Total Quality Data Management* (TdQM) (ver figura 1), apresentada por English [English 1999], procura ser uma abordagem às actividades e à actuação que devem estar na base de um sistema de garantia da qualidade da informação. O ciclo de melhoria contínua da informação tem o seu início com a avaliação dos metadados, o que inclui a definição dos dados e a sua

arquitectura. A avaliação propriamente dita dos dados dá-se no segundo processo, seguido da análise de custo imputáveis à má qualidade desses dados. Esta questão económica é importante, pois servirá para posicionar o problema utilizando a unidade monetária, algo ao qual as organizações são sensíveis, sendo também um elemento importante na análise custo/benefício de qualquer acção de melhoria. O quarto e o quinto processo destinam-se, respectivamente, às acções de melhoria da qualidade apenas da informação e à melhoria dos processos de produção/consumo de informação.

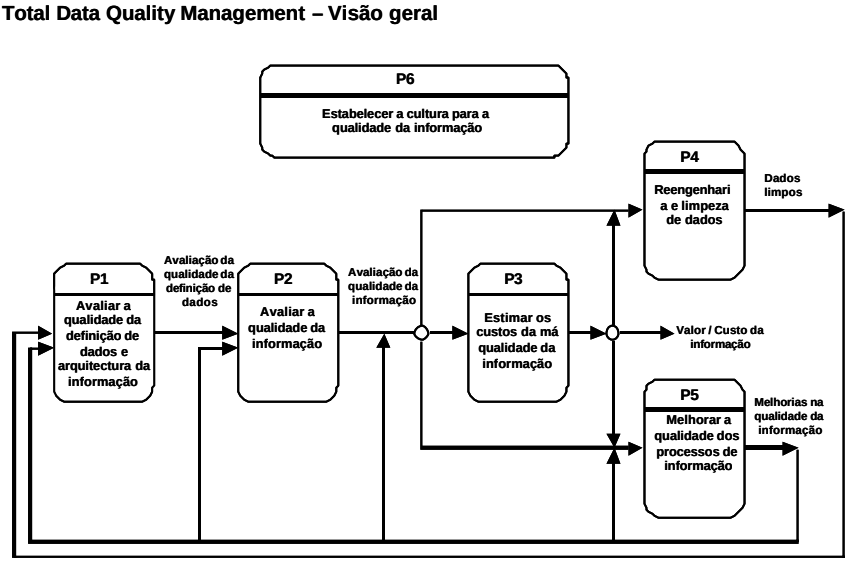

Figura 1 - Processos principais da abordagem *Total Quality Data Management* – adaptado de [Wand e R. Wang 1996]

Neste projecto particular, está-se perante um problema de migração de dados, para o qual são relevantes os processos 1, 2 e 4. Não estava em causa fazer uma análise dos custos associados à má qualidade dos dados e a melhoria dos processos estava fora do alcance da intervenção da equipa.

No que diz respeito à qualidade da definição e arquitectura dos dados, as principais deficiências que se podem encontrar são:

- Vários ficheiros que contêm informação do mesmo tipo, com definições de atributos e de domínios inconsistentes
- Inexistência de uma definição formal dos atributos ou das "*business rules*"
- A arquitectura de dados não suporta as necessidades de informação actuais

Quando as preocupações se viram para os dados em si, processo 2, aí os naturais problemas são:

- Dados em falta
- Dados incorrectos
- Duplicação de registos
- Campos utilizados com propósitos diferentes
- Valores inconsistentes entre bases de dados redundantes

Sendo um processo de migração, existe uma base de dados alvo, que precisa de ser mapeada de acordo com os dados que existem. Os problemas que aqui surgem são:

- Informação na base de dados final não existe na origem. Será possível ou mesmo apropriado atribuir valores por defeito? Será possível inferir a informação necessária?
- Há informação na origem que não é contemplada na base de dados final.
- Necessidade de limpar, traduzir e de transformar informação.

O processo de migração não se limita a definir a ligação entre o repositório de origem e o repositório final. É adequado que se faça toda uma limpeza dos dados para elevar o nível de qualidade da informação. O processo limpeza e migração dos dados pode ser estruturado do seguinte modo:

1. Identificar as fontes de dados, principalmente se há uma fonte que se sobreponha às restantes e constitua a referência para o processo.
2. Extrair e analisar os dados fonte, para verificar se os dados estão consistente com a definição conhecida deles e procurar ver relações escondidas.
3. Normalizar dados, ou seja, procurar uniformizar os conceitos que estão subjacentes aos dados, definir os atributos, domínios e *"business rules"*.
4. Corrigir, completar e uniformizar dados.
5. Anular duplicação de informação.
6. Analisar o padrão dos erros identificados, informação útil para um trabalho posterior de alteração dos processos de produzem esta informação.
7. Introduzir os dados limpos na base de dados alvo.

É de esperar que haja perdas no volume de informação final, pois alguma da informação original é descartada por não ser possível a sua correcção ou transformação. É importante reconhecer que este processo tem muito do que na industria se chama de "*scrap and rework*", algo que a qualidade procura combater. No processo de migração, o trabalho de limpeza é necessário porque todo o processo de concepção e de recolha da informação foi feito com deficiências, com lacunas de diversa ordem, e todos os dias informação com defeitos é arquitectada, recolhida, processada e armazenada.

A gestão do recurso informação vai muito mais além do que a questão tecnológica, como o software ou o hardware. À semelhança da gestão de outros recursos, é necessário aplicar técnicas de gestão que se preocupem algo mais com a informação *per si*, e que fujam de uma visão redutora da tecnologia.

Processo de migração

Como foi visto o processo de migração de dados é sempre uma tarefa de grande complexidade. No processo de migração que foi

efectuado foram equacionados previamente três cenários possíveis:

- ? Cenário 1 - Utilizar ferramentas de mercado para depuração e migração de dados;
- ? Cenário 2 - Utilizar o XML em todo o processo de migração;
- ? Cenário 3 - Desenvolver ferramentas específicas para auxílio à depuração e migração de dados.

Qualquer destes cenários genericamente teria que seguir o seguinte processo, tendo em vista o objectivo final que era a carga da base de dados da Plataforma Lattes com os *curricula vitae* (CV) mínimos dos investigadores portugueses reconhecidos pelo sistema científico nacional. O primeiro passo do processo era depurar, integrar e seleccionar a informação que iria constar na plataforma. O segundo passo consistia na geração em XML de cada um dos CV e o último passo era carregar esses CV(XML) para a base de dados relacional final já com a ligação relacional das entidades constituintes no modelo de dados do Lattes (ver figura 2).

A passagem dos CV em XML para a base de dados final da plataforma já estava implementada pelo Grupo Stela e a geração do XML também foi adaptada de sistemas já existentes, pelo que o foco da equipa portuguesa foi o de se encarregar do primeiro passo relativo à depuração e integração dos dados portugueses num base de dados com um formato previamente estabelecido mas sem as relações do modelo de dados final as quais seriam estabelecidas no sistema de carga.

Assim sendo nos cenários seguintes apenas nos iremos concentrar na primeira etapa do processo uma vez que as etapas seguintes serem iguais para todos os cenários, uma vez que efectuada a etapa inicial a qualidade dos dados já nos permitia a utilização das ferramentas existentes.

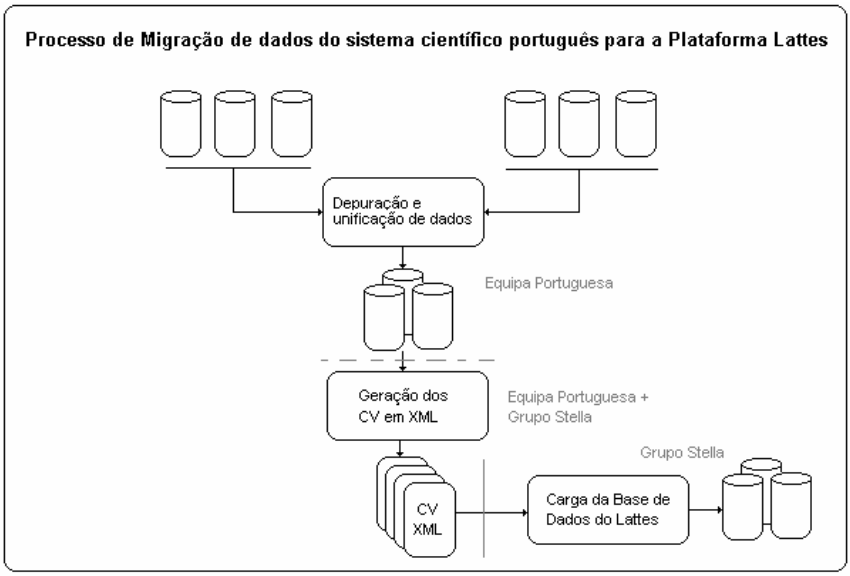

Figura 2 – Processo de migração genérico utilizado

O primeiro cenário equacionado consistia na utilização de ferramentas de mercado para a depuração e integração dos dados existentes nas várias bases de dados para cada uma das entidades. O produto desta fase seria uma base de dados única para cada entidade presente (ver figura 3).

No segundo passo aplicam-se geradores de XML para gerar os CV (*Curricula vitae*) de cada um dos investigadores, resultando após esta fase, um ficheiro XML para cada CV de investigador. Finalmente utiliza-se uma ferramenta de carga da base de dados do Lattes. A ferramenta pega em cada ficheiro XML e introduz toda a informação relativa a um CV na base de dados relacional final, criando a respectivas ligações relacionais entre todas as entidades presentes.

Este cenário apresentava alguns problemas, nomeadamente o pouco o tempo disponível para seleccionar e obter as ferramentas mais adequadas a utilizar, agravado pela diversidade de bases de dados existentes para a mesma entidade com informação

incompleta, redundante e inconsistente em alguns casos. Pelo que este cenário acabou por ser abandonado.

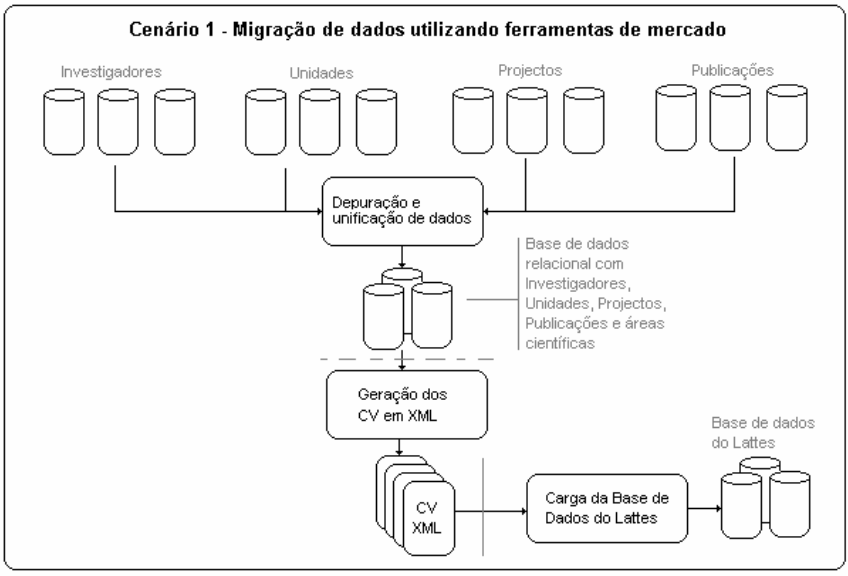

Figura 3 – Cenário 1

O segundo cenário consistia em utilizar o XML em todo o processo de migração (ver figura 4).

Neste cenário as ferramentas de geração do CV em XML tratavam da depuração e integração dos dados das bases de dados originais.

Este processo foi deixado de lado uma vez que a qualidade dos dados existentes inicialmente exigiam um processo prévio de depuração e integração que não podia de uma forma simples e eficiente ser feito juntamente com a geração de XML. Por isso este cenário também foi abandonado.

O terceiro e último cenário equacionado consistia em, por um lado desenvolver um conjunto de ferramentas de apoio à depuração dos dados para cada uma das entidades e por outro desenvolver um conjunto de ferramentas de integração dos dados das diversas bases de dados numa base de dados única com as diferentes

entidades relacionadas de modo a aplicar o gerador de XML para produzir os CV's que por sua vez seriam carregados na base de dados do Lattes (ver figura 5).

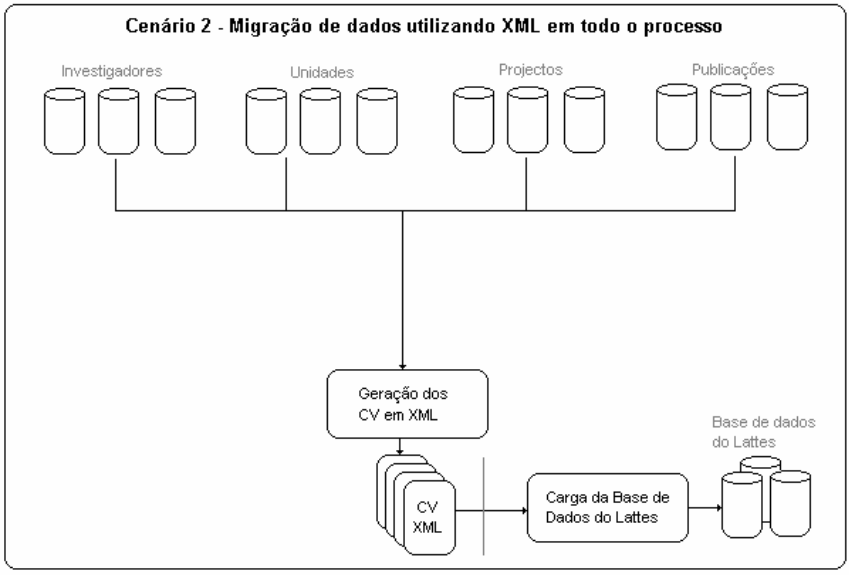

Figura 4 – Cenário 2

Este cenário foi o escolhido uma vez que era o que melhor se ajustava às características das bases de dados existentes e já anteriormente referidas.

A migração dos dados

O contexto do sistema nacional de ciência e tecnologia do Ministério da Ciência e tecnologia, em que a processo de migração dos dados foi equacionado era constituído por dois organismos, cada qual com um conjunto de bases de dados com informação relevante para a plataforma. A figura 6 as principais bases de dados utilizadas onde é também indicado o número de registos encontrados para cada uma das suas principais entidades informacionais.

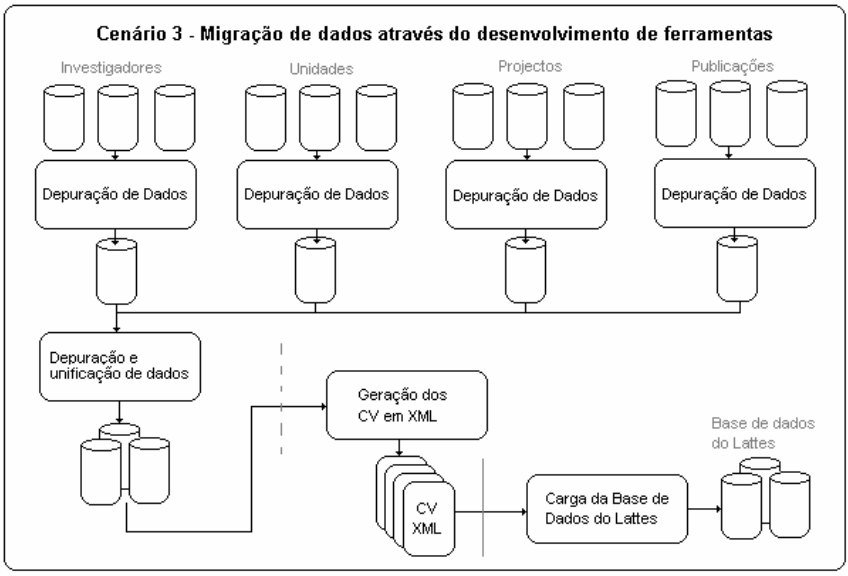

Figura 5 – Cenário 3

Assim do OCT foram utilizadas as seguintes bases de dados:

? As Bases de Dados relativas ao Inquérito sobre Potencial Científico e Tecnológica Nacional de 2000, contendo, no que para este processo é relevante, as tabelas de:

- ✍ OCTInvestigadores com 30.766 registos
- ✍ OCTUnidades com 1.469 registos

? As Bases de Dados do ISI relativas à publicação portuguesa no período compreendido entre 1981 e 2001:

- ✍ OCTISI19812001 com 34.247 registos

? As Bases de Dados do CSH relativas às publicações portuguesas na área das Ciências Sociais e Humanas, no período compreendido entre o ano de 1999 e 2000:

- ✍ OCTCSH19992000 com 22.052 registos

Da FCT foram utilizadas as seguintes bases de dados:

- As Bases de Dados do financiamento plurianual dos centros de investigação cientifica nacionais:
  - FCTPlurianualUnidades com 341 registos
  - FCTPlurianualaInvestigadores com 14.599 registos
- As Bases de Dados de projectos de investigação financiados pela FCT no ano de 1998:
  - FCTProjectos1998 com 1.748 registos
  - FCTProjectosEquipas1998 com 3.742 registos
- As Bases de Dados de projectos de investigação financiados pela FCT nos anos de 1999 e 2000:
  - FCTProjectos19992000 com 779 registos

Estas bases de dados tinham origens, finalidades e informação referente a períodos diferentes, o que trazia grandes problemas à sua integração. A agravar a esta situação há que acrescentar o facto de haver muita informação incompleta, informação referente à mesma coisa escrita de formas diferentes e entidades caracterizadas de formas diferentes conforme a Base de Dados.

Perante este contexto optou-se por seguir o cenário três, uma vez que o primeiro passo do processo era muito complexo e exigia que as ferramentas permitissem grande flexibilidade ao nível da especificação das regras de depuração e unificação. O sucesso desta tarefa era crucial para a qualidade do resultado final. As ferramentas foram construídas com Visual Basic, SQL e Access.

Neste cenário, o primeiro passo após o estudo aprofundado da arquitectura dos dados e das bases de dados originais, consistiu na

construção de ferramentas para apoio à depuração e unificação dos dados oriundos das diversas fontes.

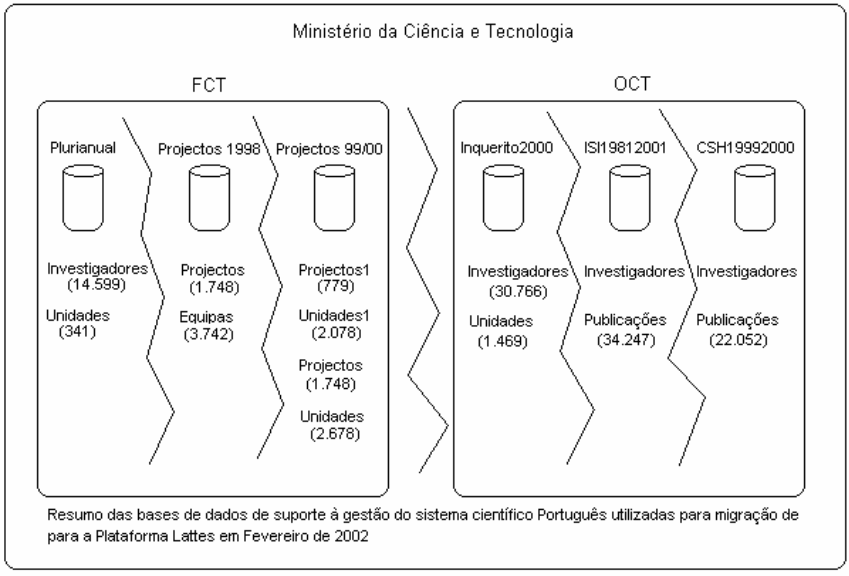

Figura 6 – Origem dos dados utilizados

O resultado desta fase foi o mapeamento de toda a informação existente nas diversas Bases de Dados numa única Base de Dados contendo apenas a informação relevante para a plataforma Lattes. No entanto, a arquitectura dos dados ainda não dispunha das relações do modelo de dados final.

Após o processo de mapeamento da informação numa única base de dados o resultado obtido foi o seguinte:

- Investigadores 28.854 registos
- Instituições 1.620 registos
- Equipas de Projectos 3.676 registos
- Projectos 2.525 registos
- Publicações 230.281 registos

Esta base de dados foi transformada em XML dando lugar a 28.854 CV dos investigadores portugueses reconhecidos pelo

sistema científico nacional. Posteriormente estes CV foram carregados para a base de dados relacional da plataforma a qual constitui o ponto de arranque do sistema em Portugal.

Conclusões

Apesar das dificuldades encontradas devido à dispersão, redundância, inconsistência e dados incompletos, a missão foi concluída com sucesso. Foi possível construir uma única base de dados segundo a arquitectura Lattes, onde cada um dos 28.854 investigadores recuperados já tem associada uma entrada com toda a informação que os antigos sistemas dispunham.

A base de dados produzida irá permitir que Portugal lance a sua versão da plataforma com os dados mais relevantes dos seus investigadores tendo estes apenas que completar os dados omissos do seu *curriculum vitae*.

Julga-se que desta forma a plataforma poderá atingir o nível de operação e utilização desejado mais rapidamente do que se tivessem seguido outras abordagens, como por exemplo lançar a plataforma sem nenhum *curriculum vitae* introduzido.

Acreditamos que desta forma se contribuiria para o lançamento em tempo oportuno da plataforma Lattes em Portugal, podendo assim, o país integrar esta rede SCienTI – Rede Internacional de Intercambio de Fontes de Informação e Conhecimento para a Gestão de Ciência, Tecnologia e Inovação.

Referências

Amaral, L., L. Santos e C. A. Bernardo, *Uma visão do Sistema Cientifico e Tecnológico Português*, I Workshop da Rede SCienTI, Florianópolis, Brasil, 2002.

Barquin, R. e H. Edelstein, *Planning and Designing the Data Warehouse*: Prentice Hall, 1997.

English, L., *Improving Data Warehouse and Business Information Quality*, John Wiley and Sons, Inc, 1999

Juran, J. e F. Gryna, *Quality Planning and Analysis*, 3 ed., 1993.

Moody, D., *Metrics for Evaluating the Quality of Entity Relationship Models*, presented at ER'98 - 17th International Conference on Conceptual Modeling, Singapore, 1998.

Moody, D. e G. Shanks, *Improving the Quality of Entity Relationship Models - Experience in Research and Practice*, ER'98 - 17th International Conference on Conceptual Modeling, Singapore, 1998.

Redman, T., *Data Quality for the Information Age*: Artech House Inc., 1996.

Wand and R. Wang, *Anchoring Data Quality Dimensions in Ontological Foundations*, Communications of the ACM, 39, 1996, pp. 86-95.

Wang, R., D. Strong, e L. Guarascio, *Data Consumers Perspectives of Data Quality - TDQM-93-12*, Total Data Quality Management Group - Massachussets Institute of Technology 1994.

Watson, R., *Data Management - Databases and Organization*, 2 ed., John Wiley & Sons, 1999.