

# Integração de Dados

Disciplina de Procedência de Dados e  
Data Warehousing

Profa. Dra. Cristina Dutra de Aguiar Ciferri  
[cdac@icmc.usp.br](mailto:cdac@icmc.usp.br)

---

# Tópicos

- Trabalho desenvolvido
  - ferramenta Reconciliador v.1
  - ferramenta Reconciliador v.2
  - Sistema Urano

---

# Ferramenta Reconciliador v.1

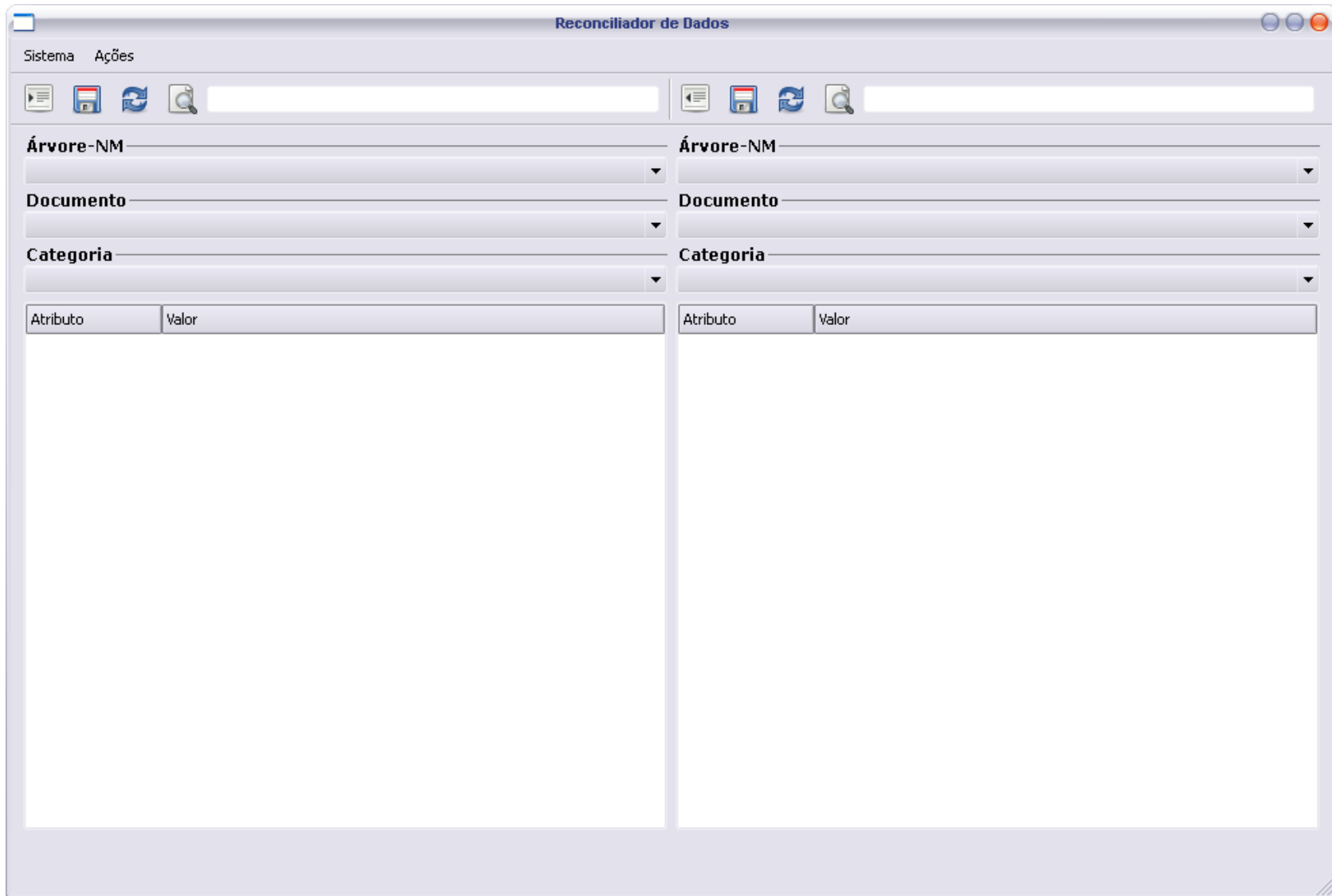
- Voltada à integração de instâncias (**conflitos de valores**)
- Características
  - semi-automatiza a identificação de
    - objetos correspondentes
    - inconsistências
  - ajuda o usuário a
    - eliminar inconsistências
    - completar dados
    - trocar dados

---

# Ferramenta Reconciliador v.1

- Aspectos funcionais
  - **visualização** dos objetos de dois documentos
    - visualização lado-a-lado
  - **sincronização** de objetos
  - **edição** de dados incompletos ou errados
  - **troca** de dados entre objetos de diferentes documentos

# Interface



# Visualização

**Reconciliador de Dados**

Sistema Ações

**Árvore-NM**  
Currículo Lattes

**Documento**  
E:\Documentos\Currículo\_Lattes\_Caetano.xml

**Categoria**  
Conference Papers

Atributo	Valor
Paper	
Title	Visually Mining an Multiple Relational Tables an Once
Year	2002
Proceedings	Proceedings of the Sixth East-European Conference ...
Paper	
Title	Classification Abstraction: an intrinsic element in Dat...
Year	2000
Proceedings	Springer-Verlag Lecture Notes on Computer Science
Paper	
Title	Enhanced Visual Evaluation of Feature Extractors for...
Year	2005
Proceedings	Proc. of the 3rd ACS/IEEE Intl. Conference on Comp...
Paper	
Title	Statistical Association Rules and Relevance Feedbac...
Year	2006
Proceedings	Prc. of the 19th IEEE Computer Based Medical Syste...
Paper	
Title	Comparing Images with Distance Functions based o...
Year	2006
Proceedings	Proceedings of the 21th ACM Symposium on Applied ...

Número de Objetos: 5  
Arquivo XML aberto.

**Árvore-NM**  
DBLP

**Documento**  
E:\Documentos\DBLP.xml

**Categoria**  
Conference Papers

Atributo	Valor
Paper	
Title	The MM-Tree: A Memory-Based Metric Tree Without ...
Year	2007
Proceedings	ADBIS
Paper	
Title	Enhanced visual evaluation of feature extractors for ...
Year	2005
Proceedings	AICCSA
Paper	
Title	Visually Mining on Multiple Relational Tables at Once
Year	2002
Proceedings	ADBIS Research Communications

Número de Objetos: 3

# Sincronização

**Reconciliador de Dados**

Sistema Ações

Árvore-NM: Currículo Lattes | Documento: E:\Documentos\Currículo\_Lattes\_Caetano.xml | Categoria: Conference Papers

Atributo	Valor
Paper	
Title	Visually Mining an Multiple Relational Tables an Once
Year	2002
Proceedings	Proceedings of the Sixth East-European Conference ...
Paper	
Title	Classification Abstraction: an intrinsic element in Dat...
Year	2000
Proceedings	Springer-Verlag Lecture Notes on Computer Science
Paper	
Title	Enhanced Visual Evaluation of Feature Extractors for...
Year	2005
Proceedings	Proc. of the 3rd ACS/IEEE Intl. Conference on Comp...
Paper	
Title	Statistical Association Rules and Relevance Feedbac...
Year	2006
Proceedings	Prc. of the 19th IEEE Computer Based Medical Syste...
Paper	
Title	Comparing Images with Distance Functions based o...
Year	2006
Proceedings	Proceedings of the 21th ACM Symposium on Applied ...

Número de Objetos: 5

Árvore-NM: DBLP | Documento: E:\Documentos\DBLP.xml | Categoria: Conference Papers

Atributo	Valor
Paper	
Title	Visually Mining on Multiple Relational Tables at Once
Year	2002
Proceedings	ADBIS Research Communications
Paper	
Paper	
Title	Enhanced visual evaluation of feature extractors for ...
Year	2005
Proceedings	AICCSA
Paper	
Paper	
Title	The MM-Tree: A Memory-Based Metric Tree Without ...
Year	2007
Proceedings	ADBIS

Número de Objetos: 3

# Edição

**Reconciliador de Dados**

Sistema Ações

Árvore-NM: Currículo Lattes | Documento: E:\Documentos\Currículo\_Lattes\_Caetano.xml | Categoria: Conference Papers

Árvore-NM: DBLP | Documento: E:\Documentos\DBLP.xml | Categoria: Conference Papers

Atributo	Valor
Paper	
Title	Visually Mining an Multiple Relational Tables an Once
Year	2002
Proceedings	Proceedings of the Sixth East-European Conference ...
Paper	
Title	Classification Abstraction: an intrinsic element in Dat...
Year	2000
Proceedings	Springer-Verlag Lecture Notes on Computer Science
Paper	
Title	Enhanced Visual Evaluation of Feature Extractors for...
Year	2005
Proceedings	Proc. of the 3rd ACS/IEEE Intl. Conference on Comp...
Paper	
Title	Statistical Association Rules and Relevance Feedbac...
Year	2006
Proceedings	Pr. of the 19th IEEE Computer Based Medical Syste...
Paper	
Title	Comparing Images with Distance Functions based o...
Year	2006
Proceedings	Proceedings of the 21th ACM Symposium on Applied ...

Número de Objetos: 5

Atributo	Valor
Paper	
Title	Visually Mining on Multiple Relational Tables at Once
Year	2002
Proceedings	Proceedings of ADBIS Research Communications
Paper	
Title	Enhanced visual evaluation of feature extractors for ...
Year	2005
Proceedings	AICCSA
Paper	
Title	The MM-Tree: A Memory-Based Metric Tree Without ...
Year	2007
Proceedings	ADBIS

Número de Objetos: 3



# Troca de Dados

The screenshot shows the 'Reconciliador de Dados' application window. The interface is split into two panes, each displaying a tree view of data objects. The left pane is titled 'Currículo Lattes' and the right pane is titled 'DBLP'. Both panes show a list of papers with their attributes (Title, Year, Proceedings) and values. The left pane shows 5 objects, and the right pane shows 3 objects. The status bar at the bottom indicates 'Sincronizado.' and 'Número de Objetos: 5' for the left pane and 'Número de Objetos: 3' for the right pane.

**Reconciliador de Dados**

Sistema Ações

**Árvore-NM**  
Currículo Lattes

**Documento**  
E:\Documentos\Currículo\_Lattes\_Caetano.xml

**Categoria**  
Conference Papers

Atributo	Valor
Paper	
Title	Visually Mining an Multiple Relational Tables an Once
Year	2002
Proceedings	Proceedings of ADBIS Research Communications
Paper	
Title	Classification Abstraction: an intrinsic element in Dat...
Year	2000
Proceedings	Springer-Verlag Lecture Notes on Computer Science
Paper	
Title	Enhanced Visual Evaluation of Feature Extractors for...
Year	2005
Proceedings	Proc. of the 3rd ACS/IEEE Intl. Conference on Comp...
Paper	
Title	Statistical Association Rules and Relevance Feedbac...
Year	2006
Proceedings	Prc. of the 19th IEEE Computer Based Medical Syste...
Paper	
Title	Comparing Images with Distance Functions based o...
Year	2006
Proceedings	Proceedings of the 21th ACM Symposium on Applied ...
Paper	
Title	The MM-Tree: A Memory-Based Metric Tree Without ...
Year	2007
Proceedings	ADBIS

Número de Objetos: 5

Sincronizado.

**Árvore-NM**  
DBLP

**Documento**  
E:\Documentos\DBLP.xml

**Categoria**  
Conference Papers

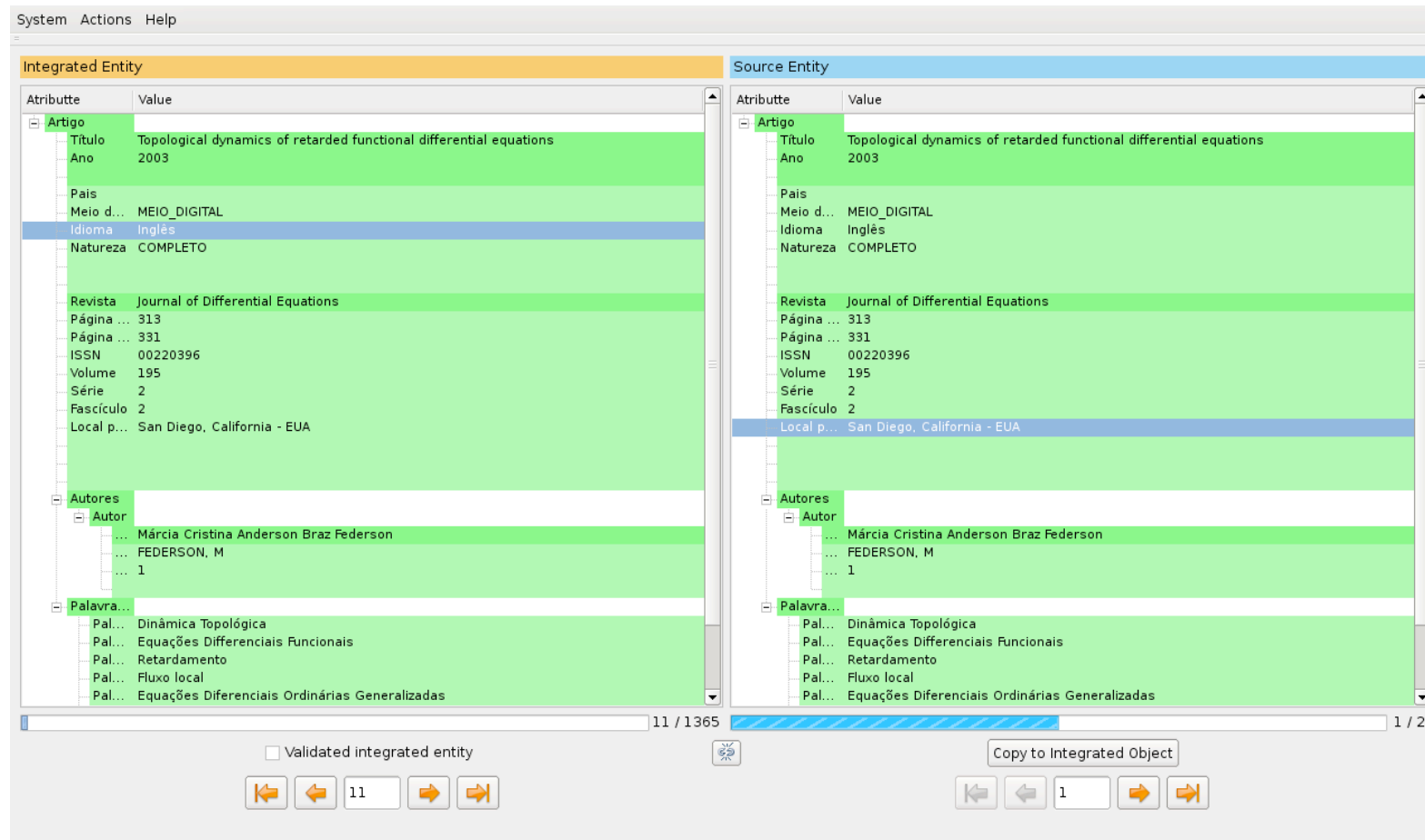
Atributo	Valor
Paper	
Title	Visually Mining on Multiple Relational Tables at Once
Year	2002
Proceedings	Proceedings of ADBIS Research Communications
Paper	
Title	Enhanced visual evaluation of feature extractors for ...
Year	2005
Proceedings	AICCSA
Paper	
Title	The MM-Tree: A Memory-Based Metric Tree Without ...
Year	2007
Proceedings	ADBIS

Número de Objetos: 3

# Ferramenta Reconciliador v.2

- Extensão da ferramenta Reconciliador v.1
- Voltada à integração de instâncias  
(**resolução de entidades**)
- Aspectos funcionais
  - geração automática de uma **entidade integrada** que melhor representa cada agrupamento
  - **mudança** de uma entidade de um agrupamento para outro agrupamento

# Ferramenta Reconciliador v.2



entidade integrada

entidades similares que pertencem  
ao mesmo agrupamento

# Geração de Entidades Integradas

- Primeira entidade-fonte
  - entidade escolhida para ser a entidade integrada é a primeira entidade-fonte pertencente ao agrupamento
- Entidade-fonte com maior peso
  - entidade escolhida para ser a entidade integrada é aquela que tem mais atributos preenchidos de acordo com pesos associados aos seus atributos
  - pesos são determinados de acordo com a importância do atributo para a entidade

# Geração de Entidades Integradas

- Entidade-fonte de maior frequência
  - entidade escolhida para ser a entidade integrada é aquela que aparece mais vezes no agrupamento
  - são considerados apenas alguns atributos da entidade para se fazer uma comparação exata entre os valores desses atributos

---

# Resultados

- Testes sobre currículos de docentes
  - entidade-fonte com maior peso produziu os resultados mais satisfatórios, desde que escolhe a entidade-fonte com maior número de atributos preenchidos
  - maioria dos agrupamentos possui apenas uma entidade-fonte (mesmos resultados para todas as propostas)
  - resultados de entidade-fonte de maior frequência foram praticamente os mesmos que os resultados de primeira entidade-fonte do agrupamento

# Mudança de Agrupamento

- Existe outro agrupamento similar à entidade-fonte
  - o usuário pode decidir mover a entidade-fonte para o agrupamento identificado pela ferramenta ou criar um novo agrupamento com a entidade-fonte em questão
- Não é encontrado outro agrupamento similar à entidade-fonte
  - o usuário pode escolher gerar um novo agrupamento para a entidade-fonte

---

# O Sistema Urano

- Dados acadêmicos dos pesquisadores
  - distribuídos em uma variedade de provedores
- Provedores heterogêneos
  - podem prover dados inconsistentes, redundantes ou complementares
- Relatórios administrativos
  - usualmente gerados manualmente
  - contêm basicamente a mesma informação, porém em formatos distintos



# Objetivos

- Coletar dados acadêmicos de provedores de informação distintos e **integrá-los** em um banco de dados centralizado
- Permitir que **relatórios** consistentes sejam gerados automaticamente

Ajuda a **tomada de decisão** e a **administração institucional**

# Provedores de Dados

- Sistemas corporativos da Universidade
  - responsável pelos **dados transacionais** relacionados às atividades de graduação e pós-graduação
  - exemplos: Sistema Júpiter e Sistema Fênix (<http://www.sistemas.usp.br>)
- Fontes externas
  - **provedores públicos** contendo dados acadêmicos
  - exemplos: Lattes (<http://lattes.cnpq.br>), DBLP

# Principais Tipos de Relatório

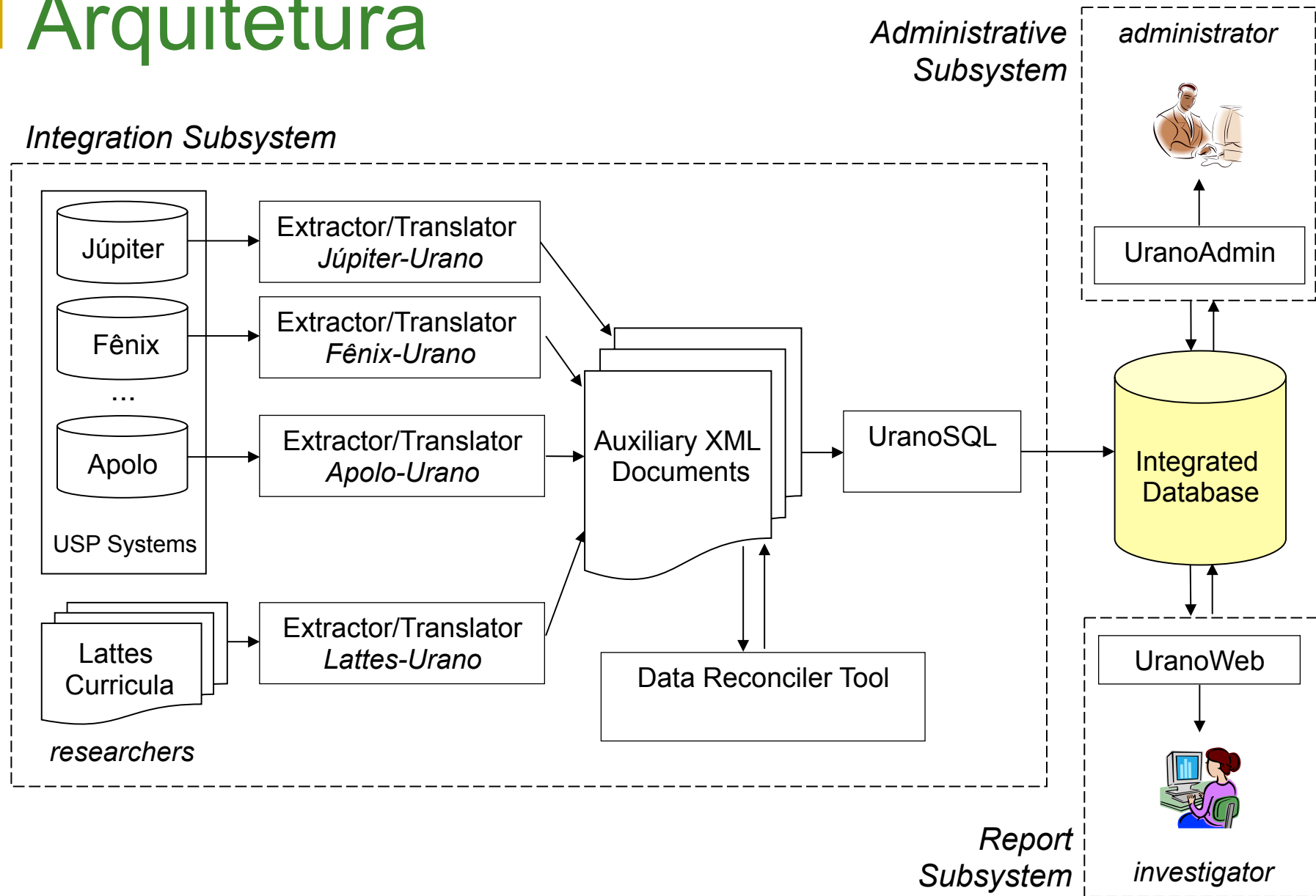
- Docentes, seus departamentos e informações profissionais
- Publicações
  - revistas, eventos, livros, capítulos de livros
- Orientações
  - iniciação científica, mestrado, doutorado, pós-doutorado
- Participação em bancas julgadoras
  - graduação, mestrado, doutorado, concurso público

---

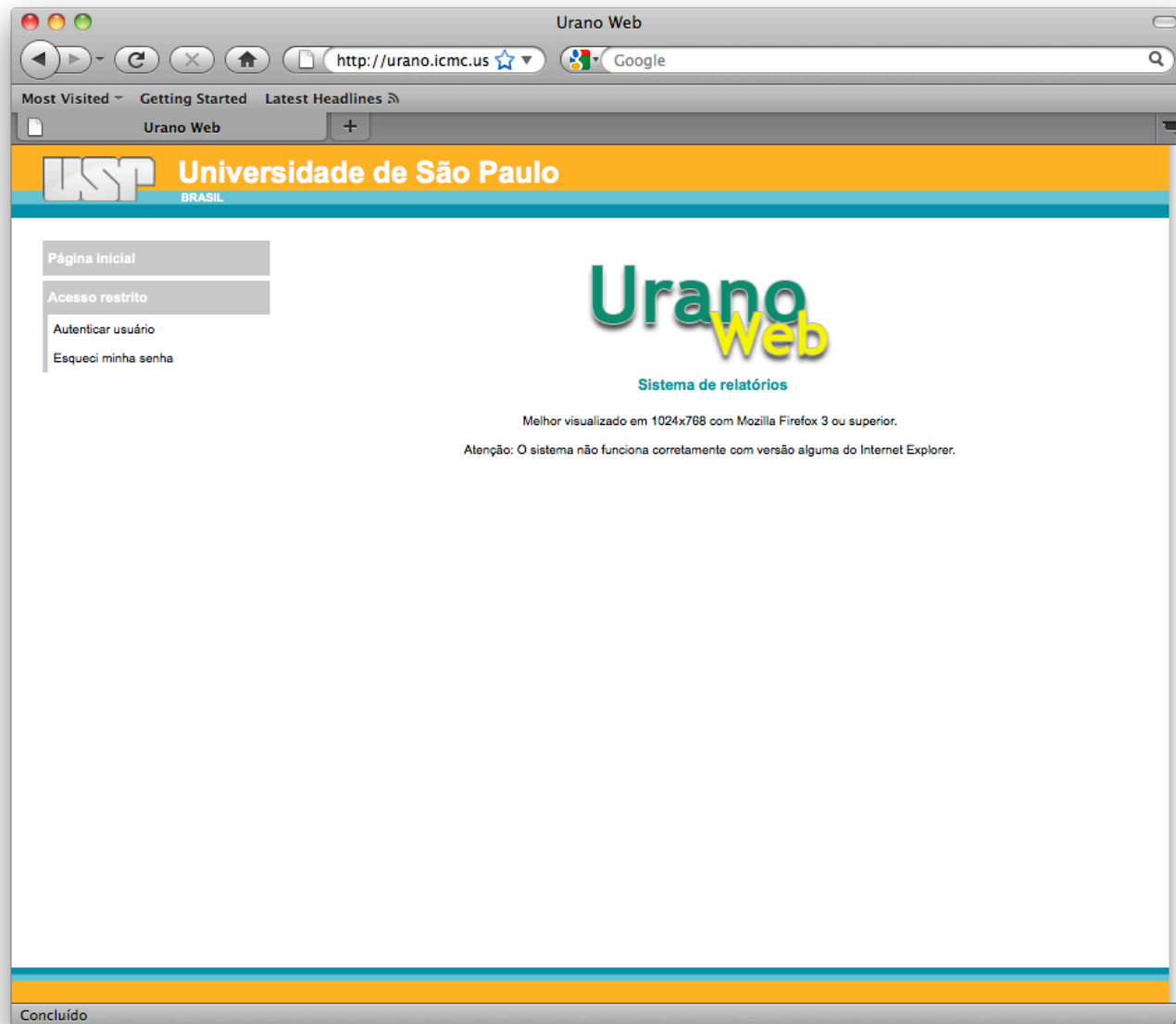
# Principais Tipos de Relatório

- Participação em conferências
- Disciplinas
  - graduação e pós-graduação
- Projetos
  - descrição, órgãos financiadores

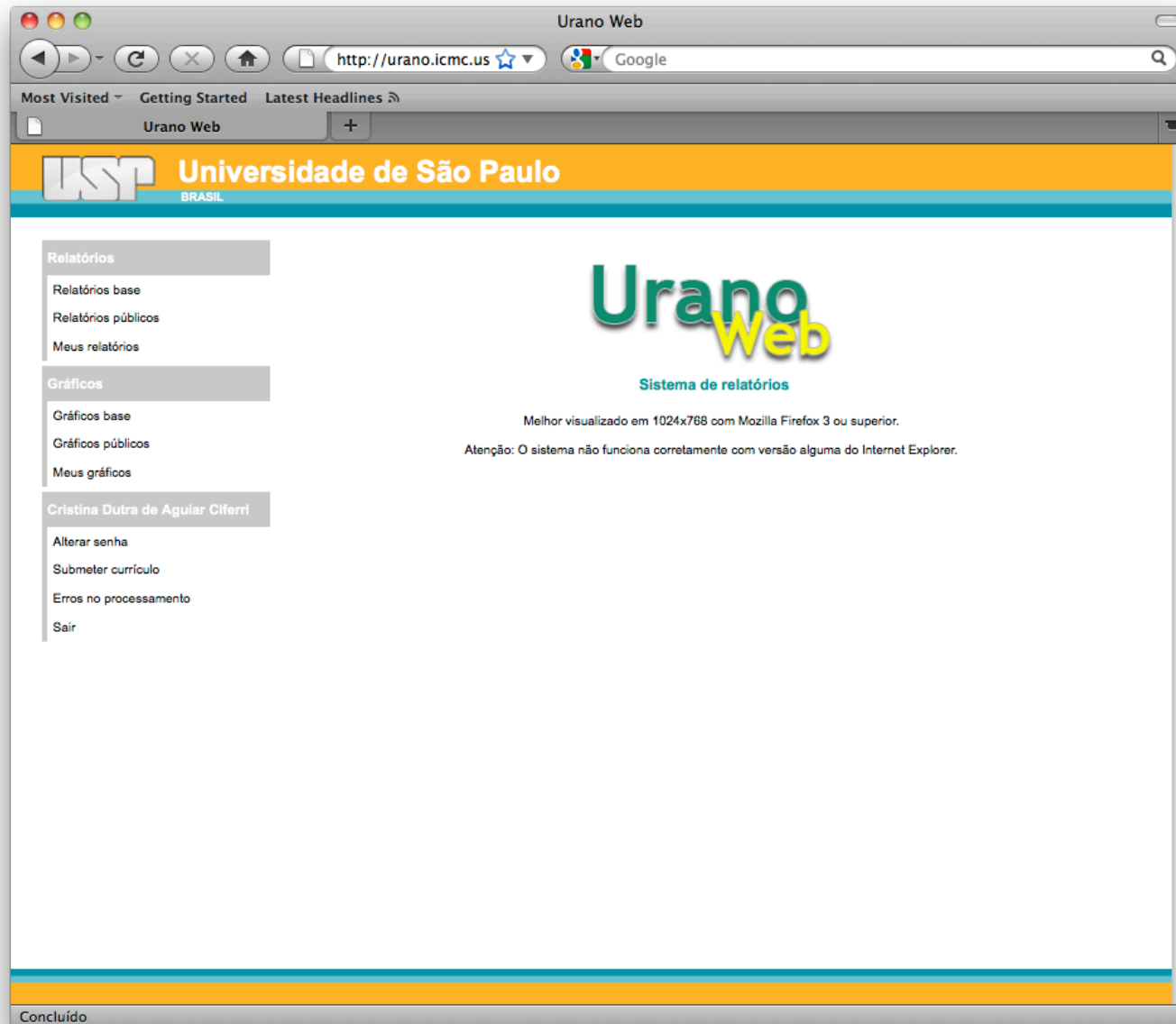
# Arquitetura



# http://urano.icmc.usp.br



# Interface



# Relatórios

The screenshot shows a web browser window titled 'Urano Web' with the URL 'http://urano.icmc.us'. The page header features the USP logo and 'Universidade de São Paulo BRASIL'. The main content area is titled 'Relatórios base' and contains a list of reports. A left sidebar provides navigation options for reports, graphics, and user actions. The bottom of the browser window shows the status 'Concluído'.

Relatório	Data
Artigos em eventos Urano Administrador	18 de junho de 2010
Artigos em revistas Urano Administrador	18 de junho de 2010
Bancas Urano Administrador	18 de junho de 2010
Capítulos de livros Urano Administrador	18 de junho de 2010
Disciplinas de graduação Urano Administrador	18 de junho de 2010
Disciplinas de pós-graduação Urano Administrador	18 de junho de 2010
Livros Urano Administrador	18 de junho de 2010
Orientações Urano Administrador	18 de junho de 2010
Participação em eventos Urano Administrador	18 de junho de 2010
Textos Urano Administrador	18 de junho de 2010
Usuários do sistema Urano Administrador	18 de junho de 2010



# Exemplo de um Relatório

The screenshot shows a web browser window titled "Urano Web" with the URL "http://urano.icmc.us". The page header features the logo of the "Universidade de São Paulo BRASIL". The main content area is titled "Artigos em revistas" and contains a "Filtros" section with the following options:

- Autor:** A text input field and a checkbox for "Sem autor".
- Enquadramento funcional:** A list of checkboxes for functional categories: MS1 - Auxiliar, MS2 - Assistente, MS3 - Doutor, MS4 - Livre-Docência, MS5 - Associado, and MS6 - Titular. Below the list are "Tudo" and "Inverter" buttons.
- Departamento:** A list of checkboxes for departments: SCC - Departamento de Ciências de Computação, SMA - Departamento de Matemática, SME - Departamento de Matemática Aplicada e Estatística, and SSC - Departamento de Sistemas de Computação. Below the list are "Tudo" and "Inverter" buttons.
- Titulo:** A text input field and a checkbox for "Sem título".
- Natureza:** A list of checkboxes for document types: Completo, Não Informado, and Resumo. Below the list are "Tudo" and "Inverter" buttons.

On the left side, there is a sidebar menu with sections for "Relatórios", "Gráficos", and user profile information for "Cristina Dutra de Aguiar Ciferri". The sidebar includes links for "Relatórios base", "Relatórios públicos", "Meus relatórios", "Gráficos base", "Gráficos públicos", "Meus gráficos", "Alterar senha", "Submeter currículo", "Erros no processamento", and "Sair". At the bottom left of the page, the word "Concluído" is displayed.

# Exemplo de um Relatório

Urano Web

http://urano.icmc.us

Most Visited ▾ Getting Started Latest Headlines ↗

Urano Web +

Ano de publicação: 2006 a 2010  
 Sem ano de publicação

Idioma de publicação:   
 Sem idioma de publicação

Pais de publicação:   
 Sem país de publicação

Situação:  
 Publicado  
 Aguardando publicação  
Tudo Inverter

**Atributos**

Atributos disponíveis	Atributos selecionados	Ano de publicação
Autor	Periódico	Rótulo: Ano de publicação
Autor em citação	Volume	Exibir: Atributo
Cidade da editora	Número	Ordem: Descendente
Departamento	Página inicial	
E-mail do autor	Página final	
Endereço lattes do autor	Ano de publicação	

**Formato**

Agrupamento: Sem agrupamento

Estilo: Referência

Tipo de referência: Artigos em revistas

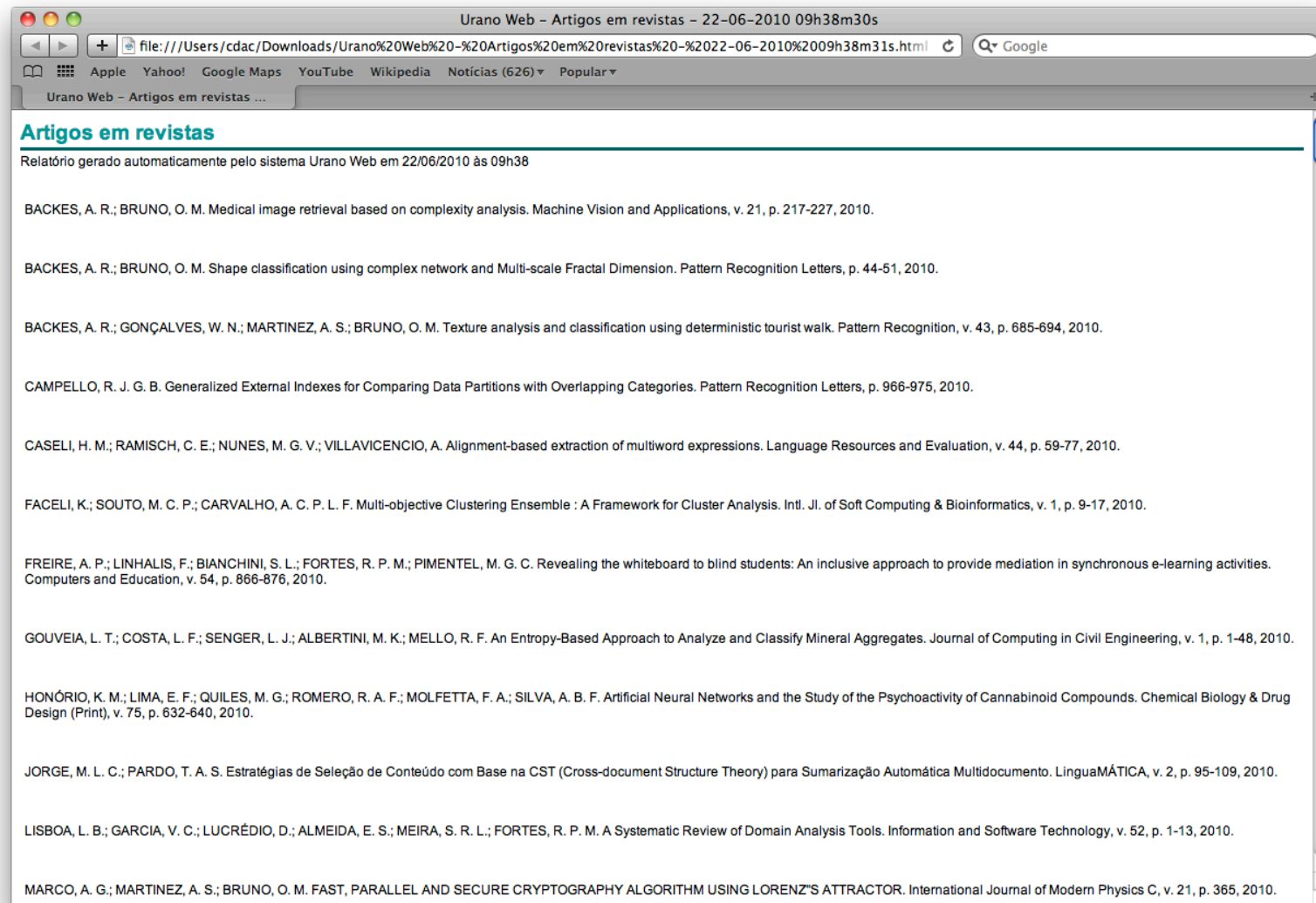
Tipo de saída: Hypertext Markup Language (HTML)

**Configuração do relatório**

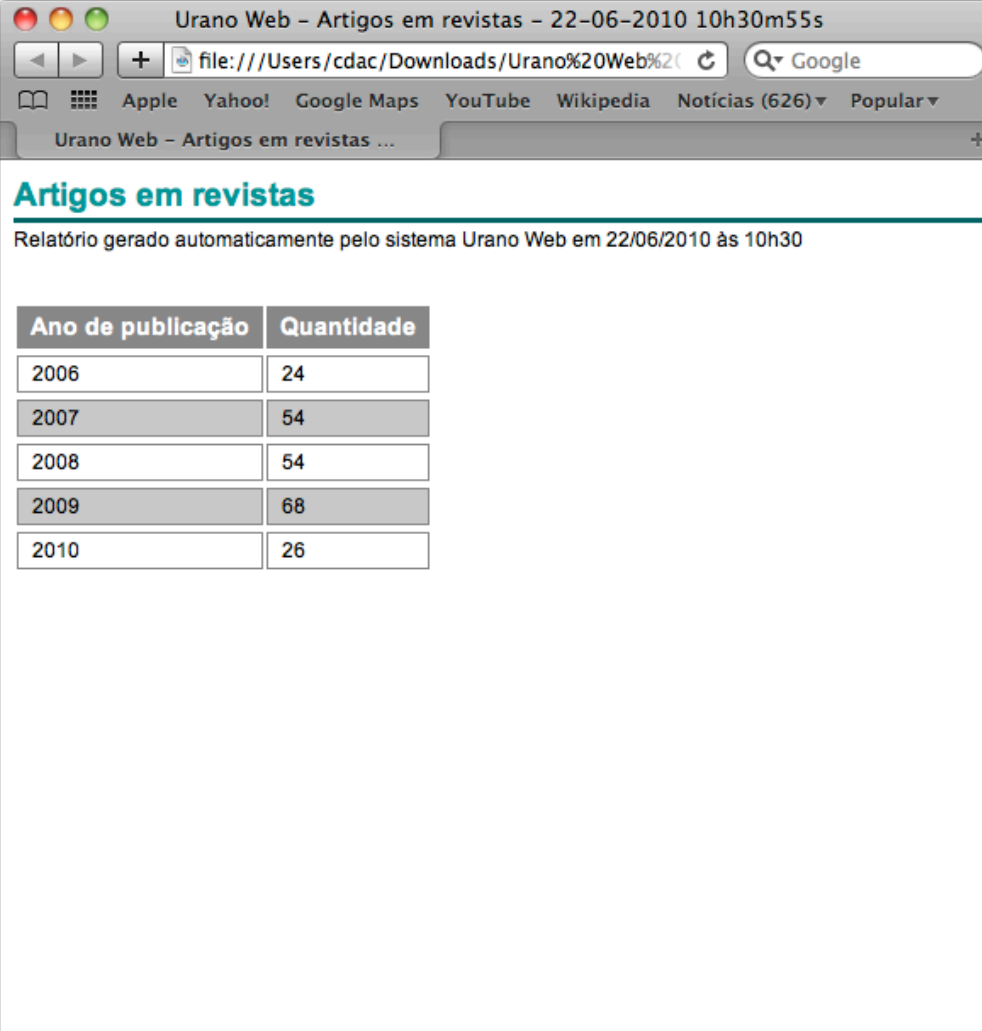
Nome: Artigos em revistas

Concluído

# Lista de Publicações em Revista



# Lista de Publicações em Revista



Urano Web - Artigos em revistas - 22-06-2010 10h30m55s

file:///Users/cdac/Downloads/Urano%20Web%20... Google

Apple Yahoo! Google Maps YouTube Wikipedia Notícias (626) Popular

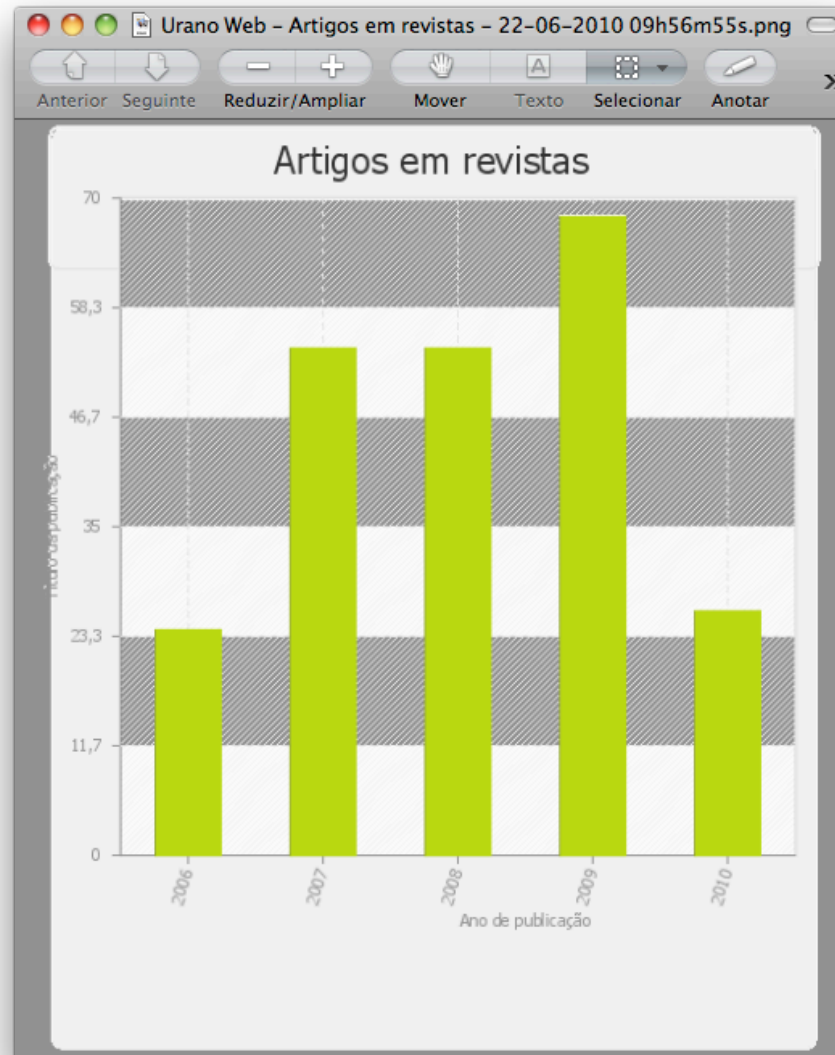
Urano Web - Artigos em revistas ...

## Artigos em revistas

Relatório gerado automaticamente pelo sistema Urano Web em 22/06/2010 às 10h30

Ano de publicação	Quantidade
2006	24
2007	54
2008	54
2009	68
2010	26

# Lista de Publicações em Revista



---

# Composição de Relatórios

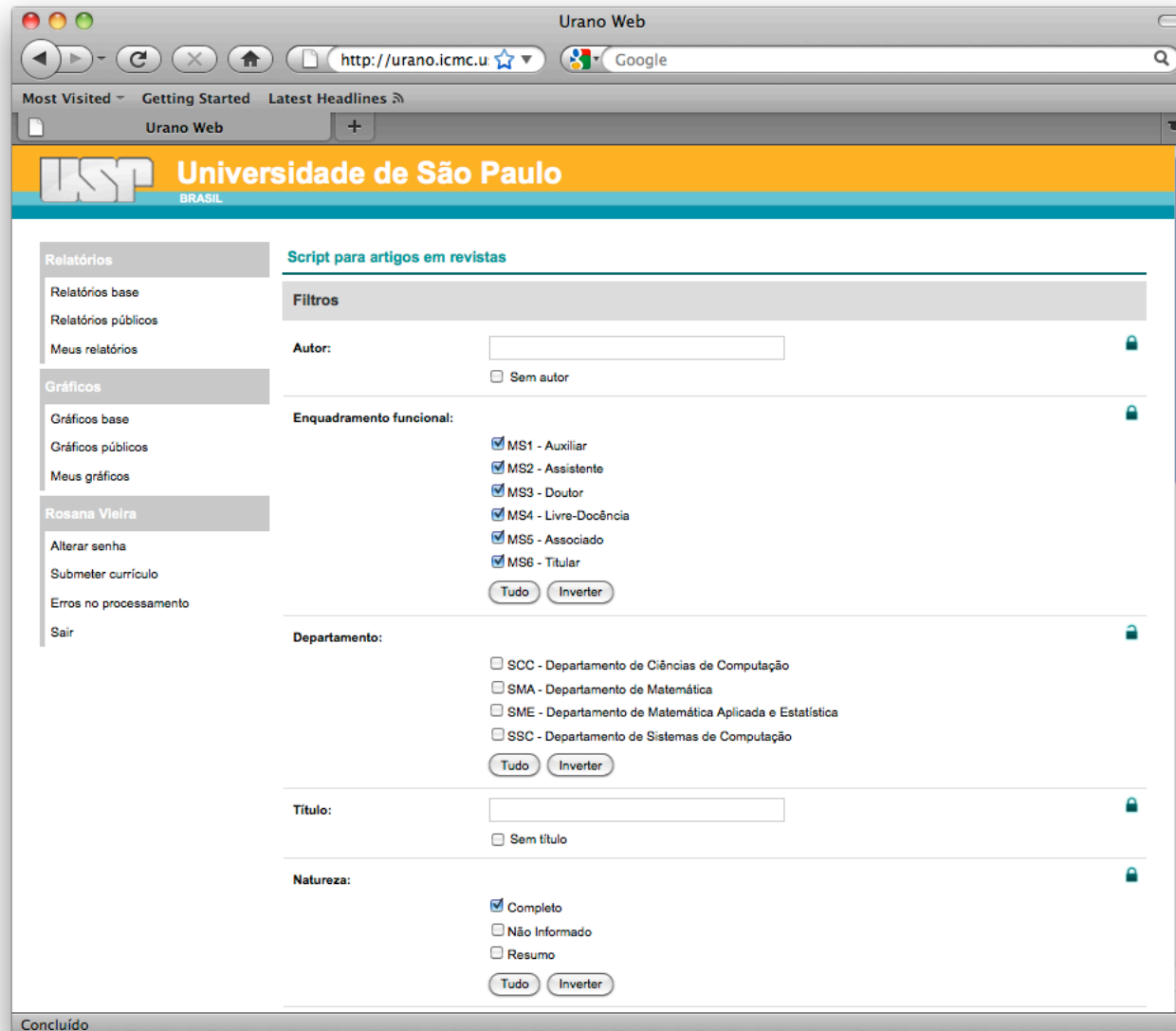
- Investigadores

- geram o formato dos relatórios que os diferentes pesquisadores devem preencher

- Pesquisadores

- preenchem o relatório de acordo com os seus dados acadêmicos particulares

# Visão do Investigador



# Visão do Investigador

Urano Web

http://urano.icmc.u... Google

Most Visited Getting Started Latest Headlines

Urano Web

Ano de publicação: 2006 a 2010  
 Sem ano de publicação

Idioma de publicação:   
 Sem idioma de publicação

País de publicação:   
 Sem país de publicação

Situação:  
 Publicado  
 Aguardando publicação  
Tudo Inverter

Tipo de publicação:  
 Artigo Científico  
 Artigo Divulgação  
 Artigo Evento  
 Capítulo  
 Livro  
Tudo Inverter

**Atributos**

Atributos disponíveis	Atributos selecionados	Lista de autores em citação
Autor	Lista de autores em citação	Rótulo: Lista de autores em citação
Autor em citação	Título da publicação	Exibir: Atributo
Cidade da editora	Periódico	Ordem: Nenhuma
Departamento	Volume	
E-mail do autor	Número	
Endereço lattes do autor	Página inicial	

**Formato**

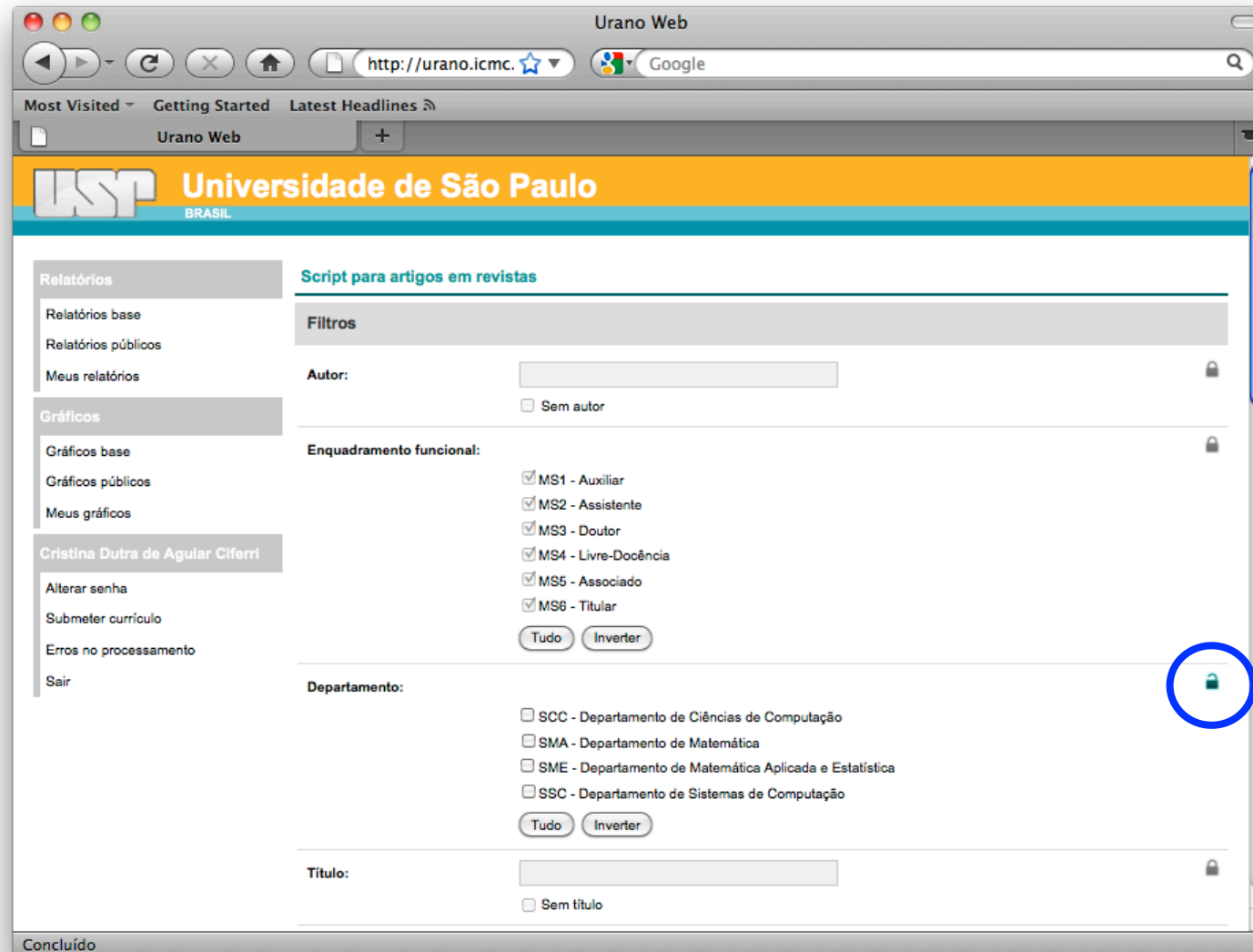
Agrupamento: Sem agrupamento

Estilo: Tabular

Concluído



# Visão do Pesquisador



# Gerenciamento do Log

The screenshot shows a web browser window titled 'Urano Web' with the URL 'http://urano.icmc.us'. The page header includes the USP logo and 'Universidade de São Paulo BRASIL'. The main content area is titled 'Erros no processamento' and contains a table comparing two records for 'Cristina Dutra de Aguiar Ciferri'.

Agma Juci Machado Traina (Fonte) X Cristina Dutra de Aguiar Ciferri (Integrado)	
<b>Dados básicos</b>	
Meio de divulgação: MEIO_MAGNETICO	Meio de divulgação: MEIO_DIGITAL
<b>Detalhamento</b>	
Cidade da editora: Porto Alegre	Cidade da editora: n?o informado
Fascículo: 1	Fascículo: n?o informado
Nome da editora: Sociedade Brasileira de Computação	Nome da editora: n?o informado
Nome do evento: Workshop on Information Visualization and Analysis in Social Networks (WIVA 2008), junto ao SBES 2008	Nome do evento: Workshop on Information Visualization and Analysis in Social Networks (WIVA 2008)
Página final: 8	Página final: 59
Página inicial: 1	Página inicial: 51
Série: 1	Série: n?o informado
Título dos anais/proceedings: Anais do WIVA 2008	Título dos anais/proceedings: Anais do Workshop on Information Visualization and Analysis in Social Networks (WIVA 2008)
Volume: 1	Volume: n?o informado

At the bottom left of the page, the word 'Concluído' is visible.