Panorama do suporte computacional às atividades de pesquisa em universidades brasileiras: um estudo de caso

Alysson Bolognesi Prado*

Maria Cecília Calani Baranauskas**

Resumo. A adoção crescente de ferramentas computacionais por cientistas em seus processos de trabalho é um fenômeno amplamente relatado, principalmente em publicações que propõem novos sistemas com esta finalidade. São escassos, entretanto, estudos que abordam numericamente esta tendência e a dimensão relativa ao total de pesquisas realizadas. Este artigo descreve um levantamento quantitativo do uso de computação em atividades de pesquisa em algumas universidades brasileiras. São apresentados, por área de conhecimento, os percentuais de uso de recursos computacionais em relação ao total de projetos de pesquisa e a evolução ao longo dos anos.

Palavras-chave. e-ciência, teses e dissertações, universidades brasileiras, suporte computacional à ciência.

An overview of computational support for research activities in Brazilian universities - a case study

Abstract. The growing adoption of computational tools by scientists in their work processes is a widely reported phenomenon, especially in publications proposing new systems for this purpose. However, studies addressing numerically this trend and the spread relative to the total surveys are less common. This article describes a quantitative survey of the use of computing in research activities of some Brazilian universities. We present, by field of knowledge, the coverage relative to the total amount of research projects as well as evolution over the years.

Keywords. e-science, theses and dissertations, Brazilian universities, computational support to science.

* Engenheiro de Computação, Doutorando em Ciência de Computação na Universidade Estadual de Campinas. E-mail: aprado@ic.unicamp.br

** Professora Titular no Instituto de Computação (IC) e Coordenadora do Núcleo de Informática Aplicada à Educação (NIED), UNICAMP, Brasil. E-mail: cecilia@ic.unicamp.br

Introdução[1]

Há algum tempo cientistas de todo o mundo têm se beneficiado de ferramentas informatizadas em suas pesquisas (Shields, 1995 *apud* Pearce, 2010; Pace *et al.*, 2010; Costa *et al.*, 2011). No início restrito a alguns poucos entusiastas, este fenômeno foi se difundindo, tornando-se parte da proposição e validação de hipóteses, bem como da simulação de modelos e da comunicação entre cientistas (Gray, 2007).

Avanços na tecnologia de redes de computadores têm permitido a agregação de equipamentos heterogêneos e geograficamente distribuídos, construindo grids[2] (FOSTER *et al.*, 2008) que somam seus poderes de processamento e armazenamento. A disponibilização desta infraestrutura geralmente é feita em um ambiente *web*, sendo gerenciado por um *middleware*[3] (APPELBE, BANNON, 2007). Dá-se o nome de e-science – ou e-ciência – ao campo multidisciplinar de pesquisa que se ocupa em propor estas novas infraestruturas para armazenamento, acesso e análise de grandes conjuntos de dados, em geral em escala acima do que pode ser manualmente gerenciado (Searight et al., 2011; Hey, Trefethen, 2002).

Para quantificar a adoção deste suporte a pesquisadores, Meyer e Schroeder (2009) estudaram as publicações disponíveis na base *Scopus*[4] através de busca por palavras-chave relacionadas a e-science. Desta análise resultaram indicadores do crescimento da adoção – principalmente a partir de 2004 – da multidisciplinaridade das pesquisas nesta área e da visão centrada nas tecnologias envolvidas na construção e operação de *grids*. Quando segmentados por área de conhecimento, os números indicam a dominância de publicações relacionadas à área de engenharia e tecnologia com 32% dos artigos, seguida da área de biológicas e saúde, com cerca de 17%.

Andronico *et al*. (2011) listaram projetos participantes da infraestrutura para e-science, que utiliza o *middleware* gLite, na Europa, América Latina, África e Ásia. Seus resultados apontam a predominância europeia de projetos em todas as áreas de conhecimento, tendo a América Latina uma presença mais significativa apenas em projetos de bioinformática. Na distribuição geral por área de conhecimento, os resultados diferem de Meyer e Schroeder (2009), pois as ciências biológicas ficam em primeiro lugar com cerca de 20% dos projetos relatados, com mais que o dobro dos projetos de engenharia, que corresponde a menos de 10%.

---

[1] Agradecimento: Este trabalho é parte do projeto EcoWeb, processo CNPq 560044/2010-0.

[2] Modelo computacional distribuído em que diversas máquinas somam suas capacidades de armazenamento e processamento de dados, de forma transparente ao usuário. Seu nome faz alusão às redes de energia elétrica domiciliares (*power grids*) que servem aos consumidores sem que estes se deem conta da infraestrutura subjacente.

[3] Programa de computador que faz a mediação entre o software executado pelo usuário e os demais elementos computacionais necessários para sua execução. Especificamente para um ambiente em grid, transporta dados e aplicativos entre os servidores, mascarando sua heterogeneidade e distribuição geográfica, quando houver.

[4] http://www.scopus.com

Trabalhos como estes, que contabilizam projetos e publicações da área, embora fundamentais para a compreensão do contexto de e-science, não verificam a proporção entre os projetos que utilizam a infraestrutura computacional e o total de projetos de pesquisa das instituições. Além disso, não existem publicações que consolidem numericamente a extensão da adoção e modalidades de uso mais frequentes de computação em pesquisa no cenário brasileiro.

O presente trabalho apresenta um levantamento quantitativo do uso de computação em nosso contexto, iniciando-se com um estudo de caso das atividades de pesquisa na Universidade Estadual de Campinas – Unicamp, a partir de análise das teses e dissertações do período 1999-2009. O estudo estendeu-se para outras universidades brasileiras, buscando ampliar o panorama desse apoio informatizado em diversas regiões do país. O artigo está organizado da seguinte maneira: na Seção 2 detalhamos os procedimentos adotados para esta pesquisa. Os resultados obtidos com o estudo de caso, através de abordagens manuais e automatizadas e sua expansão para o cenário nacional encontram-se na Seção 3. Na Seção 4 temos a discussão dos resultados e, na Seção 5, as considerações finais e perspectivas para investigações futuras.

Metodologia do Estudo

Como primeira fonte de dados, utilizamos a base de teses e dissertações disponíveis na Biblioteca Digital da Unicamp[5], cuja consulta pode ser feita manualmente através de interface web, ou automaticamente com protocolo de harvesting[6] OAI-PMH (MARTINS, SUELI, 2012; Lagose *et al.*, 2002). O universo escolhido para o estudo de caso é documentado pela publicação oficial do Anuário Estatístico de Pós Graduação (Unicamp, 2010).

Um documento foi considerado válido para este estudo quando os dados disponíveis a seu respeito na Biblioteca Digital correspondiam aos seguintes critérios:

Apresenta data de publicação registrada e dentro do intervalo de 1999 a 2009;

O título tem pelo menos 10 caracteres de comprimento;

O resumo e/ou o abstract estão disponíveis, com pelo menos 100 caracteres de comprimento;

Apresenta uma lista de palavras-chave que permite que seja associado a uma área de conhecimento.

Para uma análise exploratória do domínio, foi selecionada uma amostra aleatória de n=356 documentos cujo texto completo da tese ou dissertação está disponível em formato PDF. Esta amostra correspondente a 2% do total de itens disponíveis na biblioteca digital, para o período

[5] http://www.bibliotecadigital.unicamp.br
[6] Procedimento automatizado de coleta de dados que permite a bibliotecas digitais realizarem o intercâmbio de descritores dos seus conteúdos.

de estudo (N=17218), permitindo um erro amostral de 4% para um intervalo de confiança de 95%.

Os textos completos das teses e dissertações da amostra foram lidos em busca de referências diretas ao uso de equipamentos ou sistemas computacionais. Adotamos arbitrariamente como apoio computacional relevante aquele que ocorre em conjuntos de dados com tamanho n≥ 20; trabalhos com volumes de dados menores não foram contabilizados.

Buscou-se identificar a maneira como as ferramentas informatizadas foram utilizadas em cada uma das dissertações. Para isso foram adotadas as seguintes categorias, sendo que cada texto pode ter sido classificado em mais de uma delas:

Aquisição de dados: considera equipamentos de medição e coleta de dados conectados a computadores. Não foram considerados aparelhos digitais "standalone", ou seja, com hardware, software e interface de usuário próprios;

Armazenamento: leva em conta qualquer mecanismo de armazenamento e recuperação em Banco de Dados ou conjuntos de arquivos;

Processamento e manipulação: caracterizada pela utilização de software para tratamento de dados provenientes de experimentos, como por exemplo, pós-processamento de imagens de microscopia, filtragem de ruídos, manipulação de sequências genéticas etc.;

Simulação: uso de algoritmos e linguagens de programação para validação, extensão ou aplicação em ambiente digital de um modelo teórico;

Estatística: referência explícita ao uso de pacotes estatísticos para conjuntos de dados em que o apoio computacional é relevante;

Visualização: considera todo uso do computador como instrumento capaz de produzir inscrições gráficas do fenômeno pesquisado (Latour, 2000). Não foram considerados os gráficos gerados por pacotes do tipo Office para pequenos conjuntos (n < 20);

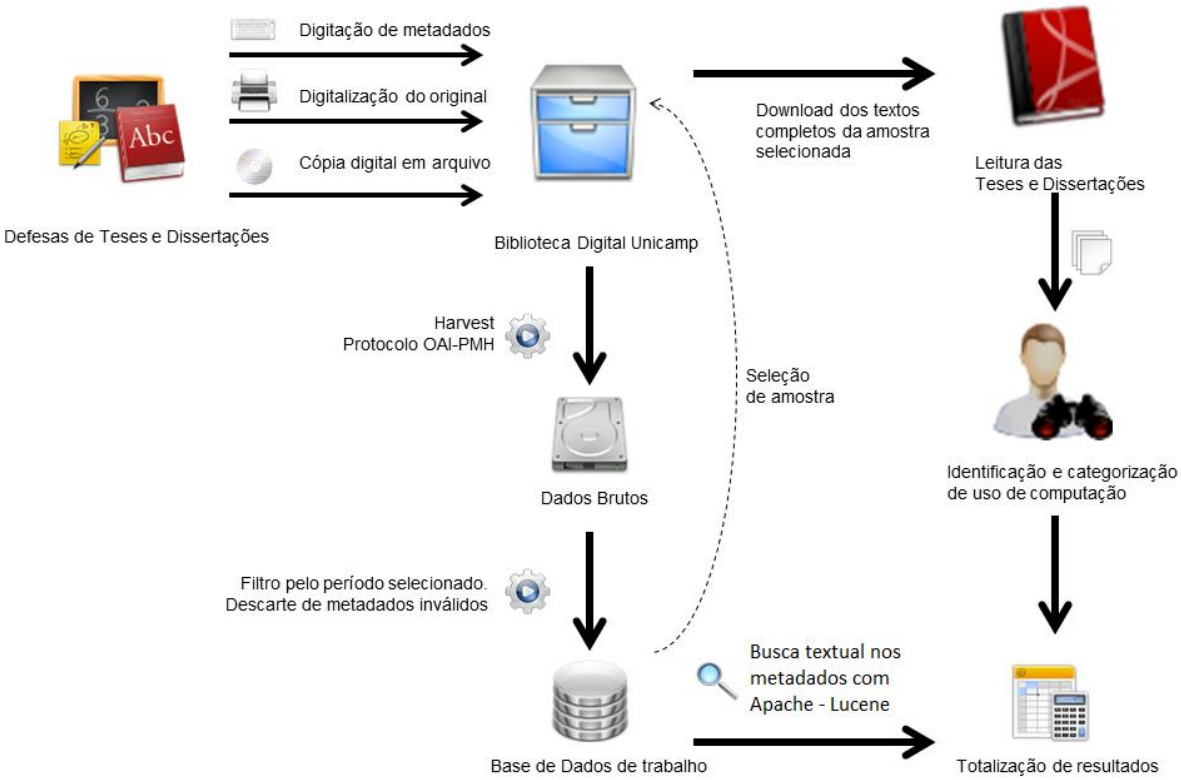
Compartilhamento: menção a qualquer produto da pesquisa que tenha sido disponibilizado em formato digital para uso por outros pesquisadores ou uso de um produto do tipo;

Objeto de estudo: casos em que a tecnologia é citada no texto não como ferramenta de apoio, mas como elemento central da pesquisa.

Para ampliar a abrangência do estudo, foi utilizado um mecanismo automatizado de busca nos metadados – mais especificamente, no *abstract* e no resumo – de palavras que indicassem o uso de ferramentas informatizadas nas pesquisas, tendo sido utilizado o software Apache Lucene[7]. Na Figura 1 estão representados os passos adotados.

[7] http://lucene.apache.org

Figura 1: Etapas e fluxo de dados no estudo de caso para os métodos manual e automatizado.



Para validação, esta abordagem foi aplicada à amostra inicial, em relação ao uso ou não de computação, realizando ajustes finos na busca de modo a minimizar a ocorrência de falsos-negativos sem potencializar o surgimento de falsos-positivos. O conjunto de palavras buscadas que demonstrou melhores resultados foi: "hardware", "software", "computador", "computer", "internet" e seus derivados – plurais, por exemplo. Obteve-se um percentual de 79,4% de confirmação na comparação entre o método manual e o automatizado, para este conjunto de palavras de busca.

Com este resultado, foi executada a busca em todos os itens do período na biblioteca digital, tendo sido obtido um total geral de 10% dos documentos com referência ao uso de computação. Considerando-se a margem de divergência entre os métodos na amostra, este número está compatível com os 27% obtidos inicialmente. Podemos então adotar este indicador automatizado como um limite inferior para o fenômeno estudado.

Para expandir o estudo para o cenário nacional, utilizamos a Biblioteca Digital de Teses e Dissertações do Ministério da Ciência e Tecnologia/IBICT[8] que concentra as produções de

[8] http://bdtd.Ibict.br

pós-graduação de quase uma centena de instituições nacionais. Foram selecionadas as entidades que dispõem de mais de 1000 publicações disponíveis e destas, aquelas com melhor qualidade nos metadados de acordo com os critérios já descritos anteriormente e, finalmente, as que apresentaram pelo menos 1% do total de documentos em cada uma das áreas de conhecimento, para que a comparação cobrisse todas as áreas.

Para determinação da área de conhecimento a que a tese ou dissertação pertence, foram usadas as seguintes abordagens, sequencialmente, até que uma delas fornecesse resultado positivo:

Algumas instituições acrescentam entre os metadados a Faculdade ou Instituto onde ocorreu a defesa. Esta informação leva diretamente à área de conhecimento. Ex: para a Unicamp e a UFMG é possível encontar "Faculdade de Engenharia ..." entre os metadados do documento;

Especificamente para a USP, a área de conhecimento nem sempre está entre os dados obtidos via *harvest*, mas consta explicitamente na interface web de sua biblioteca digital. Um pequeno programa do tipo *web crawler*[9] foi escrito para obter esta informação a partir do identificador e URL de cada arquivo;

A maioria das instituições fornece a área de conhecimento como uma palavra-chave em destaque, com os caracteres em caixa-alta. Ex: UnB, UFSCar, UFBA;

Identificou-se a lista de palavras-chave mais frequentes e genéricas que levam às áreas de conhecimento, como por exemplo "sociologia", "enfermagem" etc.

Na Figura 2 apresentamos a estrutura de dados final utilizada para armazenamento dos metadados recebidos por *harvest* e também dos resultados dos processamentos, correspondente à base de dados de trabalho. Esta estrutura permitiu acomodar os formatos de dados suportados por ambas as bibliotecas digitais. Nota-se a presença de dois campos para armazenamento do título, resumo e palavras-chave; via de regra, o primeiro contém as informações em português e o segundo, quando disponível, em inglês.

[9] Software que recupera o código HTML de uma página web a partir de seu endereço, armazenando-o localmente e permitindo extração de partes de seu conteúdo. Também permite a obtenção de recursos associados (imagens, CSS) e o acompanhamento de *hiperlinks*.

Figura 2: Estrutura de dados mapeada para a base de bados do trabalho.

```
3  public class Documento {
4
5      // dados provenientes do harvesting com OAI-DC ou MTD2-BR
6      String id;              // identificador do documento
7      String titulo1;         // título na língua original
8      String titulo2;         // título na segunda língua (inglês)
9      String url;             // endereço do documento
10     String autor;           // nome do aluno
11     String grau;            // mestrado ou doutorado (opcional)
12     String resumo1;         // resumo na língua original
13     String resumo2;         // resumo na segunda língua (inglês)
14     String dataDefesa;      // data no formato YYYYMMDD
15     String orientador;      // nome do docente
16     String instituicao;     // faculdade ou instituto de origem
17     String palavrasChave1;  // lista separada por ponto e vírgula
18     String palavrasChave2;  // idem, na segunda língua
19     String lingua;          // código da língua do conteúdo da tese
20     String setspec;         // universidade de origem
21
22     // dados resultantes dos processamentos
23     String areaConhecimento;// exatas, humanas, bio, tecn. e comput.
24     boolean analisar;       // indica que o doc faz parte do estudo
25     boolean achouPalavra;   // sucesso na busca textual?
26
27     // demais métodos e atrbutos
28     // ...
```

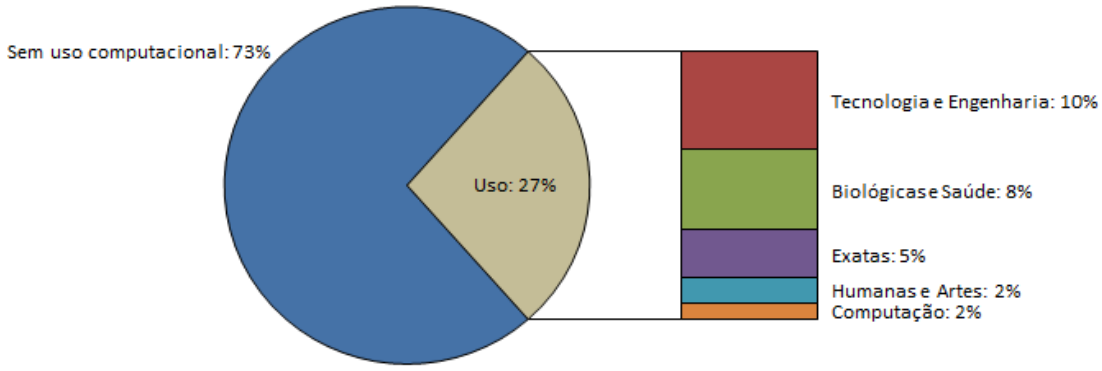
Resultados

Nesta seção apresentamos os resultados da análise detalhada das teses e dissertações da Unicamp e também das seguintes Universidades do cenário nacional: Universidade de Brasília (UnB), Universidade Federal de Santa Catarina (UFSC), Universidade Federal de São Carlos (UFSCAR), Universidade Federal de Minas Gerais (UFMG), Universidade Federal de Pernambuco (UFPE), Universidade Federal do Rio Grande do Norte (UFRN) e Universidade de São Paulo (USP).

Estudo de caso – exploração manual

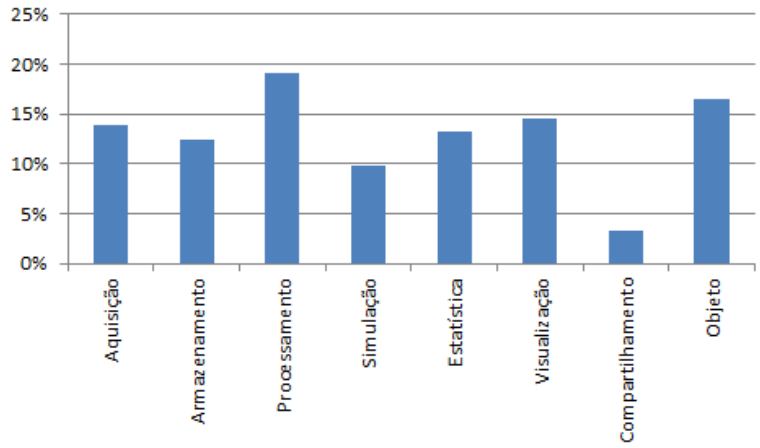
Da análise exploratória da amostra inicial das teses e dissertações da Unicamp, temos os resultados de uso de computação, em relação ao total de teses e detalhados por área de conhecimento, que estão apresentados na Figura 3.

Figura 3: Referência ao uso de computação em teses e dissertações em relação ao total da amostra.



Quanto às modalidades de uso, dos 96 textos na amostra (27% do total) que lançam mão de ferramentas informatizadas, a maior parcela (19%) se faz através de processamento e manipulação de dados, seguido de visualização e aquisição (14% cada), armazenamento e estatística (13% cada), simulação (10%) e por fim compartilhamento (cerca de 3%). O uso como objeto de pesquisa atinge 16% e estes resultados são exibidos na Figura 4.

Figura 4: Tipo de aplicação da computação, com percentual em relação ao total de casos em que foi detectado uso de computação. Um mesmo documento pode estar contabilizado em mais de uma categoria.



Estudo de caso – levantamento automatizado

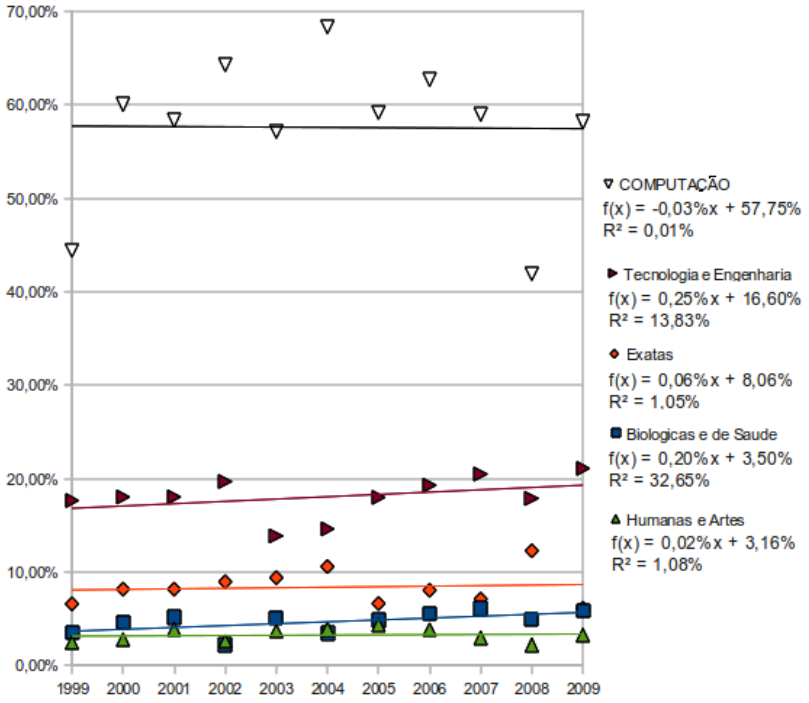
Utilizando todos os itens válidos da biblioteca digital da Unicamp (N=17218) e aplicando a abordagem automatizada descrita na seção anterior, foi possível detalhar os resultados por área de conhecimento e ano de defesa. A Tabela 1 mostra a quantidade de teses com metadados válidos disponíveis por ano, em comparação com o total oficial.

Tabela 1:Número de teses e dissertações produzidas na Unicamp e disponíveis na Biblioteca Digital, por ano.

Ano	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Anuário Estatístico	1396	1516	1908	1910	2039	2037	2066	1996	1978	1911	2092
Biblioteca Digital	1094	1200	1458	1481	1385	1592	1829	1765	1778	1740	1896
Disponibilidade (%)	78	79	76	78	68	78	89	88	90	91	91

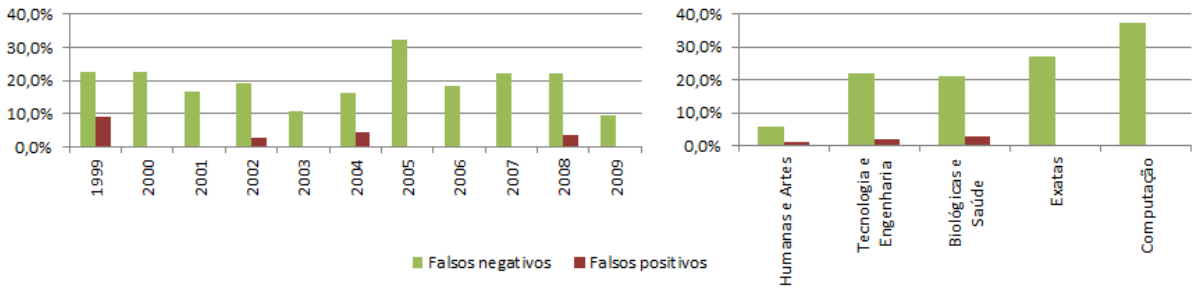
Na Figura 5 temos a evolução ao longo dos anos do percentual de teses em que foram encontradas palavras-chave relativas à computação. Nota-se que as áreas de Tecnologia e Engenharia, Biológicas e Saúde apresentam crescimento, sendo que nesta última temos os dados com comportamento mais bem ajustado ao crescimento linear. Por outro lado há estabilidade para as áreas de Exatas e Humanas, além – como esperado – da Computação. Áreas com menor valor de R^2 possuem pior ajuste à regressão linear, como é o caso de Computação, apresentando dados com mais "ruído". Analisando os dados brutos, não foi possível identificar uma tendência específica de crescimento ou queda; atribuímos os valores extremos a flutuações inerentes às limitações de um método automatizado.

Figura 5: Série temporal com o percentual de uso de computação em teses e dissertações em relação ao total de cada área de conhecimento na Unicamp. As retas no gráfico e as equações na legenda referem-se à regressão linear dos dados. R^2 é o coeficiente de ajuste da reta aos pontos.



Na comparação entre os métodos manual e automatizado, a ocorrência de falsos positivos foi esparsa e pontual (1,7% no total). Entre os falsos negativos, em relação ao ano de publicação das teses, existe uma flutuação aleatória intrínseca ao procedimento, com valor médio de 18,8% e desvio padrão de 6%. Em relação às áreas de conhecimento, o método apresenta falsos negativos em uma distribuição que pode estar relacionada à familiaridade dos pesquisadores com ferramentas computacionais em pesquisas da área. Os casos em que houve divergência são apresentados na Figura 6.

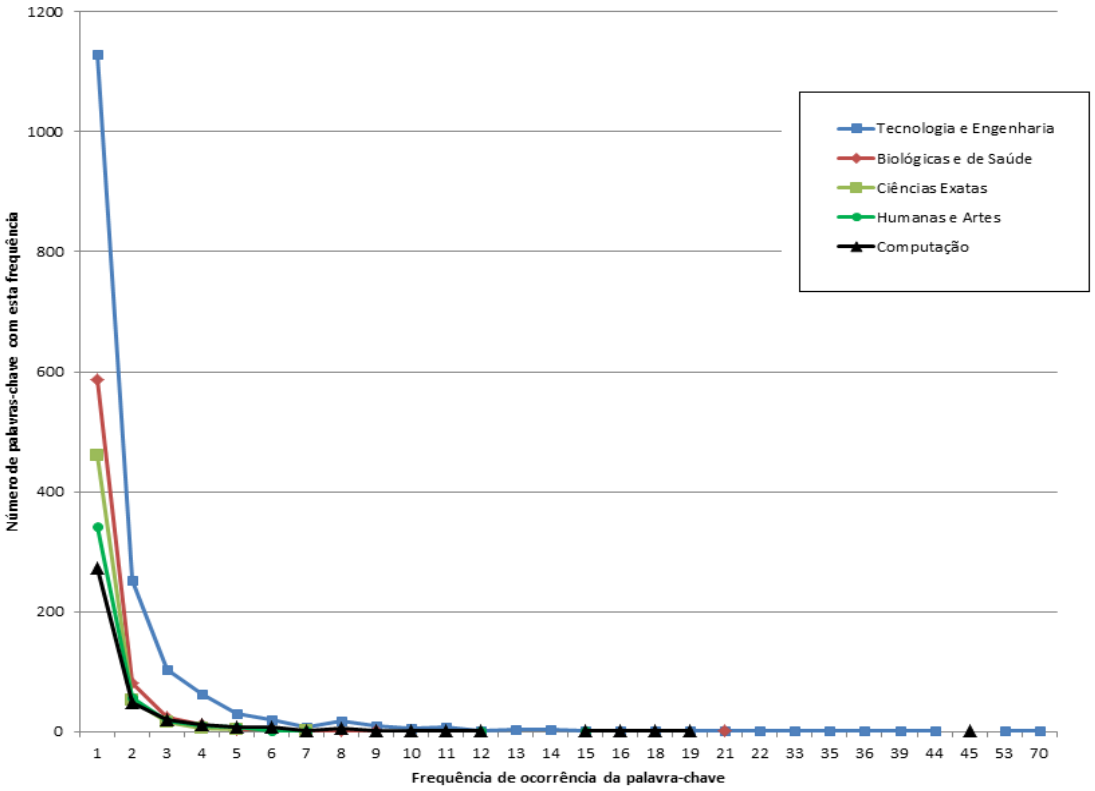
Figura 6: Divergências do método automático em comparação à leitura manual.



Os temas abordados pelas pesquisas com apoio computacional podem ser delineados pelos metadados presentes no campo "palavras chave". Na Tabela 2 estão listadas as 10 palavras-chave mais frequentes em cada área de conhecimento, dentre os documentos que apresentaram resultado positivo para a busca automatizada de uso de computação. Mesmo considerando-se que há sinônimos que não podem ser tratados automaticamente, os assuntos mais frequentes correspondem a um percentual pequeno do total de teses e dissertações com suporte computacional.

A partir da versão completa da Tabela 2, correspondente ao total de 3756 palavras-chave, foi verificada a dispersão dos assuntos que requeriram suporte computacional. Na Figura 7 estão representadas as quantidades de palavras-chave em função da frequência em que foram encontradas. Por exemplo, na área de Tecnologia e Engenharia, 1129 palavras-chave aparecem em apenas um documento; para Biológicas e Saúde, 81 palavras-chave são encontradas em dois documentos.

Figura 7: Número de palavras-chave em função do número de documentos em que aparecem, dentre as teses e dissertações com menção à computação. A maior parte das palavras-chave, indicativas do assunto abordado na respectiva tese, é encontrada em apenas um documento (à esquerda no eixo x). Temas mais frequentes, por exemplo, que aparecem em mais de 10 documentos (à direita no eixo x), são raros, encontrados cada um deles no máximo em sete textos.



.

Tabela 2: Temas mais frequentes nas teses e dissertações com apoio informatizado. O percentual é em relação ao total de documentos da área com referência a computação

Area de Conhecimento	Assunto	Nº de Teses	Percentual
Biológicas e de Saúde	Eletromiografia	21	8,0%
	Articulação temporomandibular	9	3,4%
	Materiais dentarios	8	3,1%
	Odontologia legal	8	3,1%
	Homem - Identificação	7	2,7%
	Biomecanica	6	2,3%
	Diagnostico	6	2,3%
	Endodontia	6	2,3%
	Resinas compostas	6	2,3%
	Biologia molecular	5	1,9%
Ciências Exatas	Dinamica molecular	7	3,7%
	Metodo dos elementos finitos	7	3,7%
	Industria de software	5	2,6%
	Programação não-linear	5	2,6%
	Wavelets (Matematica)	5	2,6%
	Algoritmos	4	2,1%
	Estrutura eletronica	4	2,1%
	Metodos de simulação	4	2,1%
	Otica quantica	4	2,1%
	Otimização matematica	4	2,1%
Computação	Engenharia de software	45	18,1%
	Interação homem-maquina	19	7,6%
	Software - Desenvolvimento	18	7,2%
	Tolerancia a falha (Computação)	18	7,2%
	Software - Testes	16	6,4%
	Arquitetura de computador	15	6,0%
	Otimização combinatoria	12	4,8%
	Redes de computação - Protocolos	12	4,8%
	Processamento de imagens	11	4,4%
	Redes de computação	11	4,4%
Humanas e Artes	Internet (Redes de computação)	15	10,0%
	Tecnologia educacional	12	8,0%
	Identidade	7	4,7%
	Educação a distancia	6	4,0%
	Ensino auxiliado por computador	6	4,0%
	Ambiente virtual	5	3,3%
	Analise do discurso	5	3,3%
	Educação matematica	5	3,3%
	Ensino a distancia	5	3,3%
	Informatica	5	3,3%
Tecnologia e Engenharia	Simulação (Computadores)	70	8,4%
	Redes neurais (Computação)	53	6,4%
	Modelos matematicos	44	5,3%
	Otimização matematica	39	4,7%
	Inteligencia artificial	36	4,3%
	Algoritmos geneticos	35	4,2%
	Metodo dos elementos finitos	33	4,0%
	Simulação por computador	22	2,6%
	Metodos de simulação	21	2,5%
	Otimização combinatoria	21	2,5%

Cenário com outras universidades brasileiras

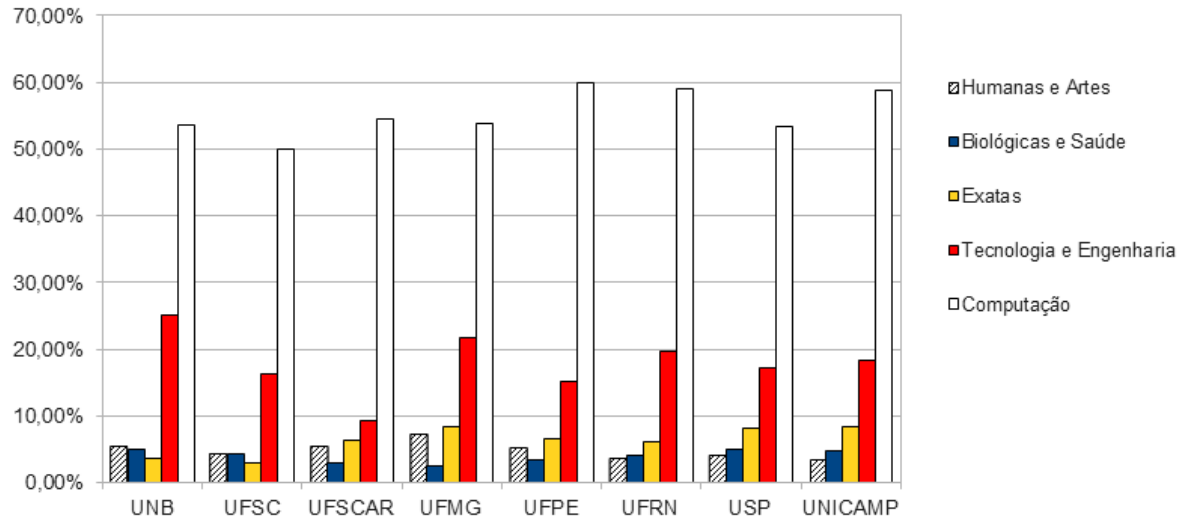
Com a base de dados proveniente da Biblioteca Digital do Ibict, efetuamos os procedimentos de limpeza, busca textual e totalização dos resultados. Na Tabela 3 apresentamos o número de teses e dissertações obtidas e válidas para análise, para as instituições selecionadas de acordo com os critérios descritos anteriormente. Além disso, a tabela contém o total de documentos analisados por área de conhecimento para o período de anos do estudo, e o número destes documentos que apresentaram resultados positivos na busca automatizada nos metadados.

Tabela 3: Total de teses e dissertações na Biblioteca Digital do Ibict, para as instituições selecionadas, documentos analisados por área de conhecimento e quantidade dos que apresentaram resultados positivos na busca automatizada.

Instituição	*Documentos*		*Analisados por área*					*Metadados positivos*				
	Disponíveis	Analisados	Hum.	Biol.	Exa.	Tec.	Comp.	Hum	Biol	Exa	Tec	Comp
UNB	5436	4543	1945	1438	624	508	28	107	71	22	127	15
UFSC	1352	1317	417	252	132	494	22	18	11	4	80	11
UFSCAR	4657	2697	753	633	448	672	191	41	19	28	62	104
UFMG	8420	3606	1195	990	216	969	236	88	25	18	210	127
UFPE	6659	6483	2636	1599	805	901	542	141	54	52	137	324
UFRN	4084	2535	799	678	328	664	66	30	27	20	130	39
USP	36295	21744	5258	9315	1981	4512	678	221	458	163	779	361
UNICAMP	38209	17218	4480	5488	2251	4575	424	150	262	190	834	249

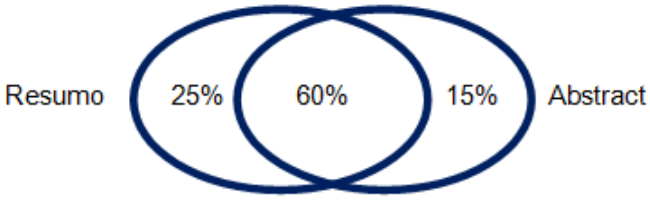
Estes dados podem ser visualizados em forma de gráfico na Figura 8. O predomínio da área de Tecnologia e Engenharia ocorre em todas as Universidades estudadas, destacando-se UFMG e UnB em relação às demais. Ciências Exatas vêm em seguida, exceto para UFSC. A área de Biológicas e Saúde é a que apresenta o comportamento mais homogêneo dentre as Universidades sendo, porém, nitidamente ultrapassada pela área de Humanas na UFMG e UFSCar. Considerando-se os índices de falsos negativos por área de conhecimento, entretanto, os perfis das instituições encontram-se bastante similares.

Figura 8: Percentual de referência ao uso de computação em teses e dissertações para algumas universidades brasileiras, em relação ao total de cada área de conhecimento.



Nesta base de dados foi verificada a necessidade de usar ambos os campos de resumo em português e em inglês – *abstract*, quando disponível – para a busca textual automatizada. Utilizando-se separadamente cada um dos campos e comparando-se com o resultado original, obteve-se a distribuição vista na Figura 9. Apenas 60% dos documentos possuem conteúdo nos dois campos com equivalência suficiente para serem detectados independentemente da língua e 25% não seriam detectados se utilizássemos apenas a versão em inglês.

Figura 9: Efetividade da busca no resumo em português versus o *abstract* em inglês.



Discussão

Conforme apontado no decorrer do artigo, ferramentas informatizadas aparecem em 27% das pesquisas descritas em teses e dissertações da Unicamp, tendo aumentado sua participação nos últimos anos, principalmente nas áreas de Ciências Biológicas e de Tecnologia e Engenharia. As outras universidades nacionais estudadas apresentaram perfis semelhantes, o que pode sugerir um padrão, a ser investigado, para a realidade brasileira.

Comparando os resultados apresentados na Seção 3 com os trabalhos relacionados apresentados na Seção 1, o perfil da pesquisa brasileira se alinha ao encontrado por Meyer e Schroeder (2009) no que diz respeito à dominância da área de Tecnologia e Engenharia. Por outro lado, tanto naquele estudo quanto no de Andronico et al. (2011), a presença de computação nas pesquisas em biologia é marcante, enquanto no cenário nacional de teses e dissertações ela ainda apresenta pouca participação, em todas as instituições estudadas, apesar de estar em crescimento.

No estudo de caso, a diversidade de uso sugere que a pesquisa em ferramentas computacionais para suporte a cientistas deve vir acompanhada de estudos teóricos e conceituais em todos os modos de utilização e não apenas no armazenamento de dados. Corrobora-se, assim, outros estudos que defendem a necessidade de agregar ao programa de e-science dimensões humanas e organizacionais como o significado dos dados, as intenções de seus produtores e também sua curadoria e confiabilidade (MEDEIROS, CAREGNATO, 2012; APPELBE, BANNON, 2007; Maeder, 2008; PEARCE, 2010). A dominância do uso para processamento aponta a necessidade do compartilhamento não apenas de dados, mas também de códigos-fonte de programas (SLIMAN et al., 2012) e *workflows* de procedimentos (MACÁRIO *et al.*, 2010).

A diversidade também é notada em relação aos temas abordados pelos trabalhos em que há apoio informatizado, selecionados pelo método automatizado. A distribuição vista na Figura 7 segue o formato de "cauda longa" – long tail (ANDERSON, 2004) – em que são poucos os assuntos frequentes e muitos os temas isolados. Por um lado, este espalhamento exprime o fato de existir benefícios com a informatização de forma razoavelmente independente do tema pesquisado; por outro, acrescenta desafios em relação a compreender e atender cada uma das especificidades (PLALE, 2012). Está fora do escopo do trabalho refinar o mecanismo de busca textual em metadados descrito na Seção 2 para apontar fielmente todas as teses que usaram sistemas informatizados. Para a finalidade pretendida – fornecer um indicador numérico automatizado cobrindo todo o conjunto de documentos – ele se mostrou suficiente. Além disso, um ajuste fino em um determinado conjunto de dados pode ter o efeito contrário em outros, inviabilizando estudos comparativos mais amplos.

Para extensão deste trabalho ao cenário internacional iniciou-se o uso da Networked Digital Library of Theses and Dissertations (NDLTD)[10], que agrega metadados de mais de uma centena de Universidades por todo o mundo. Foram selecionadas 15 instituições – cerca de

[10] http://www.ndltd.org

10% do total para um levantamento piloto, que consistiu no mesmo procedimento de *harvest* e armazenamento local dos metadados utilizado anteriormente, feitos os devidos ajustes e traduções dos termos de busca para os diversos idiomas.

De um total de 257795 documentos disponíveis, 27055 (10,5%) se mostram aptos a serem estudados de acordo com os critérios já definidos. Boa parte não dispõe do resumo com tamanho suficiente para busca, ou palavras-chave que permitam a identificação da área de conhecimento. Além disso, apenas 15% em média das teses e dissertações cujo conteúdo não é originalmente em língua inglesa dispõem de um *abstract* complementar. Conforme visto na Figura 9, isso pode aumentar significativamente a quantidade de falsos-negativos e inviabilizar a comparação com os estudos anteriores. Desta forma, apesar da grande disponibilidade de dados, a diversidade de línguas se mostra um grande obstáculo.

Conclusão

É inegável no mundo contemporâneo o papel que a tecnologia computacional tem desempenhado em todos os setores da atividade humana. Embora a atividade científica seja inerentemente a que maior benefício poderia ter da tecnologia computacional, a dimensão desse uso em pesquisa nas diferentes áreas do conhecimento ainda é pouco conhecida; esse conhecimento é fundamental para se caracterizar e localizar oportunidades de desenvolvimento em e-science.

Os números mostram que a adoção de ferramentas computacionais é significativa em relação ao total de pesquisas, que em algumas áreas de conhecimento está em crescimento – como em Biológicas e Tecnologia – e que há um perfil similar entre as instituições estudadas. Este estudo contribui para um maior conhecimento sobre a diversidade de usos e necessidades em instituições de pesquisa brasileiras, direcionando esforços de pesquisa de ferramentas computacionais de suporte à ciência.

Do ponto de vista metodológico, a experiência adquirida mostra que compreender quantitativamente a adoção de ferramentas informatizadas por cientistas, a partir do seu registro textual em teses e dissertações, pode ser feito em contextos específicos, em que se dispõe de informações complementares, na forma de metadados detalhados e confiáveis.

Artigo recebido em 06/02/2013 e aprovado em 22/03/2013.

Referências

ANDERSON, C. The long tail. *Wired Magazine,* n. 12, Oct. 2004.

ANDRONICO, G. et al. E-infrastructures for e-science: a global view. *Journal of Grid Computing,* n. 9, p.155-184, 2011.

APPELBE, B.; BANNON, D. E-research: paradigm shift or propaganda?. *Journal of Research and Practice in Information Technology,* v. 39, n. 2, May 2007.

COSTA, R. et al. E-science-as-a-service: desafios e oportunidades para a criação de nuvens científicas. In: CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, 31., 2011. *Anais...* [S.l.: s.n.], 2011.

FOSTER, I. et al. Cloud computing and grid computing 360-degree. In: IEEE GRID COMPUTING ENVIRONMENTS (GCE08), 2008. *Proceedings...* [S.l.: s.n.], 2008.

GRAY, J. *E-science:* a transformed scientific method. [S.l.]: Microsoft Research, 2007. Prefácio do livro The Fourth Paradigm.

HEY, A.; TREFETHEN, A. The UK e-science core programme and the grid. *Future Generation Computer Systems,* v. 18, n. 8, p. 1017-1031, 2002.

LAGOZE, C. et al. *The Open Archives Initiative protocol for metadata harvesting:* technical specification version 2.0. [S.l.: s.n.], 2002.

LATOUR, B. *Ciência em ação.* São Paulo: Editora da Unesp, 2000.

MACÁRIO, C.; SOUSA, S.; MEDEIROS, C. Play it again, SAM: using scientific workflows to drive the generation of semantic annotations. In: IEEE INTERNATIONAL CONFERENCE ON E–SCIENCE, 6., 2010. *Proceedings*... [S.l.: s.n.], 2010.

MAEDER, A. E-research meets e-health. In: AUSTRALASIAN WORKSHOP ON HEALTH DATA AND KNOWLEDGE MANAGEMENT, 2., 2008, Austrália. *Proceedings*... [S.l.: s.n.], 2008.

MARTINS, D.; SUELI, S. Protocolo OAI-PMH e sistemas federados de informação. *Liinc em Revista,* v. 8, n. 2, 2012.

MEDEIROS, J.; CAREGNATO, S. Compartilhamento de dados e e-science: explorando um novo conceito para comunicação científica. *Liinc em Revista,* v. 8, n. 2, 2012.

MEYER, E.; Schroeder, R. Untangling the web of e-research: towards a sociology of online knowledge. *Journal of Infometrics,* n. 3, p. 246-260, 2009.

PACE, T,; BARDZELL, S.; FOX, G. Practice-centered e-science: a practice turn perspective on Cyberinfrastructure design. In: ACM INTERNATIONAL CONFERENCE ON SUPPORTING GROUP WORK - GROUP'10, 16., 2010. *Proceedings*... [S.l.: s.n.], 2010.

PEARCE, N. A study of technology adoption by researchers. *Information, Communication & Society,* v. 13, n. 8, p. 1191-1206, 2010.

PLALE, B. Managing the long tail of science: data and communities. In: CONFERENCE OF THE EXTREME SCIENCE AND ENGINEERING DISCOVERY ENVIRONMENT, 1., 2012, Chicago. *Proceedings*... Chicago, 2012.

SEARIGHT, H. et al. E-research in the social sciences: the possibilities and the reality. *Current Research Journal of Social Sciences,* v. 3, n. 2, p. 71-80, 2011.

SHIELDS, M. The legitimation of academic computing in the 1980s. *Work and Technology in Higher Education,* p. 161-187, 1995.

SLIMAN, L.; CHARROUX B.; STROPPA Y. RunMyCode: an innovative platform for social production and evaluation of scientific research. In: IEEE INTERNATIONAL CONFERENCE ON COMPUTATIONAL ASPECTS OF SOCIAL NETWORKS (CASON), 4., 2012. *Proceedings*... [S.l.: s.n.], 2012.

UNIVERSIDADE DE CAMPINAS - UNICAMP. *Anuário estatístico da pós-graduação 2009.* Campinas: Universidade Estadual de Campinas, 2010.