UNIVERSIDADE FEDERAL DE MATO GROSSO
INSTITUTO DE FÍSICA
PROGRAMA DE PÓS-GRADUAÇÃO EM FÍSICA AMBIENTAL

INFRAESTRUTURAS VIRTUAIS DE COMPUTAÇÃO CIENTÍFICA APLICADAS NA MITIGAÇÃO DE HETEROGENEIDADE SEMÂNTICA DE DADOS AMBIENTAIS

RODICRISLLER RODRIGUES

JOSIEL MAIMONE DE FIGUEIREDO

Brasil

2015

UNIVERSIDADE FEDERAL DE MATO GROSSO
INSTITUTO DE FÍSICA
PROGRAMA DE PÓS-GRADUAÇÃO EM FÍSICA AMBIENTAL

INFRAESTRUTURAS VIRTUAIS DE COMPUTAÇÃO CIENTÍFICA APLICADAS NA MITIGAÇÃO DE HETEROGENEIDADE SEMÂNTICA DE DADOS AMBIENTAIS

RODICRISLLER RODRIGUES

Dissertação apresentada ao Programa de Pós-Graduação em Física Ambiental da Universidade Federal de Mato Grosso, como parte dos requisitos para obtenção do título de Mestre em Física Ambiental.

JOSIEL MAIMONE DE FIGUEIREDO

Brasil
2015

Dados Internacionais de Catalogação na Fonte.

R696i Rodrigues, Rodicrisller.
Infraestruturas Virtuais de Computação Científica Aplicadas na Mitigação de Heterogeneidade Semântica de Dados Ambientais / Rodicrisller Rodrigues. -- 2015
104 f. : il. color. ; 30 cm.

Orientador: Josiel Maimone de Figueiredo.
Dissertação (mestrado) - Universidade Federal de Mato Grosso, Instituto de Física, Programa de Pós-Graduação em Física Ambiental, Cuiabá, 2015.
Inclui bibliografia.

1. Computação Científica. 2. e-Science. 3. Heterogeneidade Semântica. 4. HDI. 5. D4Science. I. Título.

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

UNIVERSIDADE FEDERAL DE MATO GROSSO
INSTITUTO DE FÍSICA
Programa de Pós-Graduação em Física Ambiental

FOLHA DE APROVAÇÃO

TÍTULO: INFRAESTRUTURAS VIRTUAIS DE COMPUTAÇÃO CIENTÍFICA APLICADAS NA MITIGAÇÃO DE HETEROGENEIDADE SEMÂNTICA DE DADOS AMBIENTAIS

AUTOR: RODICRISLLER RODRIGUES

Dissertação de Mestrado defendida e aprovada em 30 de setembro de 2015, pela comissão julgadora:

Prof. Dr. Josiel Maimone de Figueiredo
Orientador
Instituto de Computação – UFMT

Prof. Dr. Allan Gonçalves de Oliveira
Examinador Interno
Instituto de Computação – UFMT

Prof. Dr. João Paulo Delgado Preti
Examinador Externo
Instituto Federal de Mato Grosso - IFMT

A Deus,
à Bem-Aventurada Virgem Maria e
aos meus pais.

Agradecimentos

À Santíssima Trindade, Pai, Filho e Espírito Santo. À Ela toda honra e toda glória pelos séculos dos séculos.

À Santíssima Virgem que em Sua maternal consolação tudo dispôs, segundo a vontade de Seu Filho, para que esse mestrado se realizasse.

À minha família, de modo especial aos meus pais e irmãos, por todo amor e solicitude com que me ajudaram e apoiaram em todas as fases desse mestrado e da vida.

A todos meus estimados amigos pessoais, que não cito nominalmente para evitar injustiças, pela compreensão, afeto e ajuda nas diversas etapas desse empreendimento.

Ao professor Josiel Maimone de Figueiredo, que foi meu grande colaborador nesse trabalho, pela dedicação, compromisso e zelo que orientou esse projeto; pelo animo e apoio dados nos momentos mais difíceis e conturbados que a sua realização enfrentou. Também pelo aprendizado, experiência e companheirismo transmitidos desde o começo de minha graduação, ao me iniciar no mundo da pesquisa científica nos diversos projetos em que fui orientado por ele.

Aos professores José de Souza Nogueira, José Holanda Campelo Júnior, Allan Gonçalves de Oliveira, Geraldo Lúcio Diniz, Marta Cristina de Jesus Albuquerque Nogueira, Carlo Ralph de Musis, Marcelo Sacardi Biudes, Iramaia Jorge Cabral de Paulo, Denilton Carlos Gaio e Paulo Henrique Zanella de Arruda, pelas aulas e solicitude constante.

Aos meus amigos de sala de aula Armando da Silva Filho, Denes Martins de Morais, Guilherme Falcão da Silva Campos, Heloisa Oliveira Marques, Ivan Tocantins, Magdiel Josias do Prado, Mauro Sérgio de França e Rafael da Silva Palácios, pelos estudos, trabalhos, apresentações, projetos e demais atividades acadêmicas. Sem eles, definitivamente, não teria chegado a termo esse fase formativa.

"Muitos se enganaram por quererem parecer sábios antes do tempo, pois com isto envergonharam-se de aprender dos demais o que ignoravam. Tu, porém meu filho, aprende de todos de boa vontade aquilo que desconheces. Serás mais sábio do que todos, se quiseres aprender de todos. Nenhuma ciência, portanto, tenhas como vil, porque toda ciência é boa. [...] O bom estudante deve ser humilde e manso, inteiramente alheio aos cuidados do mundo e às tentações dos prazeres, e solícito em aprender de boa vontade de todos. Nunca presuma de sua ciência; não queira parecer douto, mas sê-lo; busque os ditos dos sábios, e procure ardentemente ter sempre os seus vultos diante dos olhos da mente, como um espelho."
(Hugo de São Vítor, Opúsculo sobre o modo de aprender)

Resumo

RODRIGUES, Rodicrisller. **Infraestruturas Virtuais de Computação Científica Aplicadas na Mitigação de Heterogeneidade Semântica de Dados Ambientais.** 2015. 102 f. Dissertação (Mestrado em Física Ambiental) - Instituto de Física, Universidade Federal de Mato Grosso, Cuiabá, 2015.

Com a difusão dos dispositivos eletrônicos e a facilidade na troca de informações qualquer projeto científico usa dados oriundos de diversas fontes. Nesse contexto, dados científicos possuem aspectos semânticos inerentes ao seu processo de aquisição e área de domínio, de forma que para trabalhar com conjuntos de dados de origens diferentes é preciso resolver o problema da heterogeneidade semântica entre eles. Este trabalho enfoca a mitigação da heterogeneidade em dados científicos através do uso de infraestruturas virtuais de Computação Científica. O objetivo principal é demonstrar que soluções em nuvem computacional possuem os requisitos necessários para atender às demandas atuais de um contexto de Computação Científica e *e-Science*, de forma a facilitar a troca e interação de dados entre contextos heterogêneos

Palavras-chave: Computação Científica. e-Science. Heterogeneidade Semântica. HDI. D4Science.

Abstract

RODRIGUES, Rodicrisller. **Virtual Infrastructure Applied on Scientific Computing Mitigation of Semantic Heterogeneity of Environmental Data.** 2015. 102 f. Dissertação (Mestrado em Física Ambiental) - Instituto de Física, Universidade Federal de Mato Grosso, Cuiabá, 2015.

With the diffusion of electronic devices and the facility in exchange of information any scientific project uses data from different sources. In this context, scientific data have semantic aspects inherent to the process of acquisition and domain area, in such a way to work with different sources of data sets is necessary to solve the problem of semantic heterogeneity between them. This work focuses on the mitigation of the heterogeneity on scientific data using virtual infrastructures for Scientific Computing. The main objective is to demonstrate that computational cloud solutions have the necessary requirements to meet the current demands of the context of Scientific Computing and e-Science, in order to facilitate the exchange and interaction of data between heterogeneous environments.

Keywords: Scientific Computing. e-Science. Semantic Heterogenity. HDI. D4Science.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

API	Application Programming Interface
BPM	Business Process Management
CERN	Conseil Européen pour la Recherche Nucléaire
CRISP-DM	Cross Industry Standard Process for Data Mining
ENIAC	Electronic Numerical Integrator Analyzer and Computer
ETo	Evapotranspiração de Referência
FAO	Food and Agriculture Organization
HDI	Hybrid Data Infrastructures
IaaS	Infrastructure as a Service
IaaS	Infrastructure as a Service
INMET	Instituto Nacional de Meteorologia
LHC	Large Hadron Collider
LOFAR	Low Frequency Array
NASA	National Aeronautics and Space Administration
OMM	Organização Mundial de Meteorologia
OWL	Web Ontology Language
PaaS	Platform as a Service
PGFA	Programa de Pós-Graduação de Física Ambiental
SaaS	Software as a Service
SEE-GRID	South-East European GRid e-Infrastructure Development
SGBDs	Sistemas Gerenciadores de Bancos de Dados
SIGs	Sistemas de Informações Geográficas
SQL	Structured Query Language
SSM	Summary Schemas Model

StatMan	Statistical Manager
TabMan	Tabular Data Manager
UFMT	Universidade Federal de Mato Grosso
USC	University of Southern California
VO	Virtual Organization
VRE	Virtual Research Environments
Web	World Wide Web

Sumário

1 Introdução

Desde os primórdios da Computação até os dias atuais, ela sempre se apresentou como uma ferramenta de grande utilidade para comunidade científica nas mais interessantes e complexas atividades que o homem realiza. Porém, foi nas últimas décadas que a Computação, em seu amplo aspecto de técnicas algorítmicas, infraestrutura e ciência, passou de ferramenta útil para a resolução de problemas para o status de quase imprescindível. Ao ponto de ela ser considerada nos dias atuais o terceiro pilar metodológico da ciência, ao lado da teoria e da experiência. Isso aconteceu devido ao aumento acelerado da quantidade de dados armazenados, procedente de sensores e todo tipo de aparelhos eletrônicos de mensuração. Eles estão cada vez mais precisos e numerosos, o que faz com que as informações geradas por eles sejam impossíveis de serem analisadas manualmente.

O uso da Computação no contexto de pesquisa científica recebeu o nome de Computação Científica, e mais recentemente de *e-Science*. É nesse contexto que, juntamente com o aumento dos dados coletados, cresceu a percepção de como esses dados possuíam divergências de significado entre si. Isso ocorre pois, dependendo da área do conhecimento ou da abordagem teórica escolhida pelo pesquisador, os dados passam a ser descritos e observados por apenas um recorte específico. Essa escolha metodológica acaba se refletindo na forma como esses dados são armazenados, processados e compartilhados. Essa diferença na forma de interpretar os dados coletados é denominada *heterogeneidade semântica*. É com esse tema que este trabalho se ocupará.

De forma mais ampla, quando houve a primeira explosão do número de bases de dados nos anos 80, esse problema já foi sentido em um cenário que ficou conhecido como “ilhas de informação”. Com o aparecimento da Web nos anos 90, e a expansão da rede para áreas além da militar e científica, a situação se agravou mais um pouco. No âmbito científico o estudo sobre o assunto levou ao esclarecimento de algumas distinções teórica. Entre elas se localizou o problema de heterogeneidade semântica quando há o desenvolvimento temporal do significado dos dados dentro de um domínio científico ou quando há um distribuição espacial dos conjuntos de dados.

Por essa natureza ampla, a heterogeneidade semântica pode estar presente em vários níveis ou camadas das técnicas de Computação Científica. Nesse sentido, como um recorte mais específico para esse trabalho foi escolhido a heterogeneidade semântica presente em nível de dados e suas consequências para a integração deles.

As consequências da heterogeneidade semântica variam de acordo com o nível de dificuldade de comunicação entre os conjuntos de dados. Elas vão desde a necessidade da aplicação de transformações simples até um processamento mais complexo. Esses trata-

mentos mais complexos são realizados, quando possíveis, para se extrair de determinado conjunto de dados as informações que são relevantes para aquele domínio. Saber tratar essas situações é crucial para a capacidade da ciência de agregar cada vez mais informações a fim de ter um objeto de estudo quantitativamente mais bem descrito, de modo que possibilite novas descobertas naquela área de conhecimento.

No Programa de Pós-Graduação de Física Ambiental (PGFA) podem ser identificados várias manifestações do problema de heterogeneidade semântica e suas consequências. Isso ocorre, por exemplo, no processo de aquisição de dados brutos dos diversos tipos de sensores. Muitas vezes os dados adquiridos pelos pesquisadores são provenientes de diversas fontes, que não mantém entre si uma homogeneidade no tratamento das informações coletadas. Uma outra ocorrência se dá no processamento desses dados. Ainda que se considere uma situação em que as fontes de dados forneçam as informações sem heterogeneidade semântica na coleta, o processamento desses dados por modelos e análise estatísticas muitas vezes é realizado por planilhas eletrônicas ou algoritmos primitivos. Essas técnicas não seguem uma padronização semântica na saída dos dados para posterior armazenamento e compartilhamento. Essas e outras situações, acrescidas da própria evolução do domínio na compreensão dos valores coletados, fazem com que esses dados semanticamente heterogêneos sejam de difícil agregação à outros conjuntos de dados provenientes de coleta ou processamento. Essa homogeneidade semântica que se busca é necessária, por exemplo, para formação de *Big Data* e o compartilhamento dos resultados de pesquisa já acalçados.

Como esse é um problema que alcançou não somente o meio científico com advento da Web, a busca de soluções para ele extrapolou a academia e passou a conjugar esforços de várias áreas de atividades humanas. As diversas soluções, tecnologias e abordagens que surgiram desse cenário mais amplo serviram de base para o tratamento desses problemas no âmbito científico. É nesse sentido, que as infraestruturas virtuais baseadas em Computação Distribuída e em Nuvem aparecem como uma proposta arrojada para *e-Science* para o tratamento da heterogeneidade dos dados. Nesse contexto, o presente trabalho buscou realizar uma abordagem de modo a verificar e explorar a capacidade da infraestrutura virtual *D4Science* em mitigar a heterogeneidade semântica de dados presente em atividades comuns no PGFA.

1.1 Justificativa

O recorte temático escolhido para esta dissertação é o que abrange os problemas de heterogeneidade semântica relacionados à prospecção, processamento e armazenamento de dados ambientais, no contexto do PGFA da Universidade Federal de Mato Grosso (UFMT). Um pesquisador dessa área, normalmente, lida com dados oriundos de várias fontes. Muitas vezes, essas informações são de tipos diferentes, mas que podem ser transformadas para que

significam algo dentro do domínio estudado por ele. É notável, também, o contato que esse tipo de pesquisador mantém com as várias áreas do conhecimento. O próprio Programa já é de carácter multidisciplinar, agregando estudantes das mais diversas formações. Todo esse contexto faz desse ambiente um bom exemplo de como a ciência em geral atualmente convive com as mais heterogêneas medições e abordagens teóricas.

Na pesquisa científica percebe-se a adoção dos dois componentes fundamentais da ciência, o objeto e o método com o qual ele será analisado. Na Física Ambiental, os objetos de pesquisa são, de maneira geral, os fenômenos associados às trocas de energia e matéria entre a biosfera e a atmosfera. Na análise desses fenômenos entram em cena as diversas metodologias que buscam modelar de modo lógico-quantitativo o ambiente observado.

Desde as simples variáveis escalares, passando pelas equações de estimativa de fluxos de energia, em suas mais diversas manifestações, até os modelos de saldo de radiação e balanço de energia, é notada uma variedade de elementos conceituais sendo associados a medições empíricas. Tanto na linguagem adotada para expressar os conceitos, como nos termos ligados às medições e respectivas unidades de medidas é possível ver uma heterogeneidade semântica. Ela pode ser fruto tão somente da adoção de termos que os lógicos chamam de equívocos, ou também de termos que realmente representam coisas análogas, mas que só podem ser comparadas após a devida manipulação matemática. Nesse processo, são os conjuntos de dados que apresentam o retrato mais claro da grande heterogeneidade semântica que acontece no processo da produção científica. Essas variações são normais e lícitas se tratadas de forma rigorosa, mas podem propiciar certa dificuldade para pessoas que não estão a par dos métodos e medições utilizadas por determinada pesquisa e tem acesso aos dados coletados para realização de outro trabalho.

Devido a grande relevância para o contexto do processo de produção científica na Física Ambiental é que se buscou abordar esse tema e os meios de mitigar as consequências dessa diversidade semântica. Um Programa como o PGFA que lida constantemente com grandes quantidades de dados micrometeorológicos vindos dos mais diferentes lugares e servindo aos propósitos diversos de suas linhas de pesquisa necessita de uma esforço para o esclarecimento dessa questão. A adoção de técnicas computacionais se torna imperativa para a produção científica e a capacidade dessas tecnologias de lidar com a heterogeneidade semântica deve ser levada em conta para a possível adoção de uma futura plataforma de *e-Science* para a instituição.

1.2 Objetivo Geral

O objetivo geral desse trabalho é verificar se uma infraestrutura virtual de Computação Científica é capaz de fornecer os meios para mitigar a heterogeneidade semântica em dados ambientais.

1.3 Objetivos Específicos

Com o intuito de alcançar o objetivo geral dessa pesquisa é necessário chegar ao termo de objetivos específicos que podem ser elencados como segue:

a) revisar a literatura científica a respeito do fenômeno de heterogeneidade semântica no contexto de produção científica e computacional;

b) revisar a literatura científica sobre Computação Científica e suas iniciativas que lidem com dados semanticamente heterogêneos;

c) desenvolver um estudo de caso, que contemple atividades comuns de pesquisa científica em Física Ambiental, para avaliar a capacidade de resposta de infraestrutura virtuais em lidar com os problemas de diversidade semântica;

d) identificar a presença dos tipos de heterogeneidade semântica encontrados nas situações propostas no Estudo de Caso. Em especial, os destacados pela literatura, como heterogeneidade semântica de: tempo, identificadores, granularidade, conflito de nomenclatura e unidades de medida;

e) mitigar ou tratar totalmente a heterogeneidade semântica dos conjuntos de dados selecionados nas situações propostas no Estudo de Caso;

f) testar a homogeneidade semântica dos resultados obtidos após o uso das técnicas computacionais escolhidas para mitigar a heterogeneidade semântica;

g) discutir e avaliar a capacidade da infraestrutura virtual escolhida de mitigar os problemas comuns de heterogeneidade semântica encontrados nas pesquisas em Física Ambiental.

1.4 Estrutura do trabalho

Esta dissertação segue uma estrutura na disposição de seu conteúdo, ela pode ser descrita como segue:

- **Capítulo 2:** em um levantamento bibliográfico são apresentados dois assuntos relacionados ao problema e que foram considerados necessários para o desenvolvimento do trabalho. São eles: a Heterogeneidade Semântica e a Computação Científica. As informações obtidas nessa etapa formaram o critério de execução de todas as outras;

- **Capítulo 3:** são apresentados os materiais que foram utilizados para se chegar ao objetivo deste trabalho. Aqui estão descritas as técnicas computacionais que foram adotadas, os critérios pelos quais elas foram escolhidas e também as escolhas metodológicas do trabalho;

- **Capítulo 4:** são descritas de forma detalhada todas as etapas realizadas no Estudo de Caso. Desde a avaliação dos conjuntos de dados em sua forma original até o teste de integração semântica depois que eles foram tratados pela infraestrutura virtual;
- **Capítulo 5:** são tecidas algumas considerações gerais sobre o trabalho feito e alguns apontamentos sobre o possível desenvolvimento futuro dessa pesquisa.

***

2 Fundamentação Teórica

Buscando aprofundar os assuntos de interesse do recorte temático deste trabalho, foram selecionados alguns referencias teóricos de maior relevância no estudo deles [1]. O aprofundamento nesses assuntos é o fundamento para os critérios adotados na escolha das abordagens para mitigação da diversidade semântica. Por isso, neste capítulo foi empreendido um esforço de trazer sumariamente o desenvolvimento histórico e teórico desses assuntos, porém sem negligenciar os detalhes necessários à uma explicação abrangente da questão.

As seções abaixo obedecem a seguinte distribuição. Em um primeiro momento (seção 2.1), traçou-se uma descrição do grande ambiente da Computação Científica, que compartilha áreas de intersecções com as mais diversas ciências. Entre essas ciências de contato com a Computação Científica, as que estão mais envolvidas são as que enfocam a natureza, como é o caso da Física Ambiental. Nesse espaço, buscou-se fazer um esclarecimento a respeito do uso da Computação como metodologia científica. Além disso, colocou-se em evidência que a ciência auxiliada pela Computação é um ambiente de origem da heterogeneidade semântica nos dados científicos e ambientais. Porém, a Computação Científica também foi observada como cenário de soluções para superar ou mitigar essas inconsistências semânticas. O segundo assunto abordado é o próprio conceito de heterogeneidade semântica (seção 2.2). É investigada a sua natureza, os seus tipos e as possíveis soluções para esse problema. É importante notar que nesse segundo ponto se explorou o tema da diversidade semântica em si mesmo. Os contextos específicos que foram utilizados, apenas foram citados para auxiliar na descoberta de algo a mais sobre esse conceito. Por fim, tecem-se algumas considerações sobre o trabalho realizado neste capítulo.

2.1 Computação Científica

A relação entre os computadores e a atividade científica é de longa data. Desde o *Electronic Numerical Integrator Analyzer and Computer* (ENIAC), o primeiro computador digital, até as realizações mais recentes do *Large Hadron Collider* (LHC) do *Conseil Européen pour la Recherche Nucléaire* (CERN), a Computação, em seu amplo aspecto de técnicas algorítmicas, infraestrutura e ciência, sempre deu suporte as atividades científicas mais complexas e importantes que o homem vem desenvolvendo (RUTTIMANN, 2006). Mas, recentemente, devido ao aumento vertiginoso da quantidade de dados disponíveis, provenientes de equipamentos de observação e mensuração, cada vez mais precisos e

[1] A forma e os critérios pelos quais eles foram selecionados estão descritos no Capítulo 3 na seção 3.2.

numerosos, o uso da Computação passou de mero auxílio à solução de alguns problemas para se tornar quase indispensável. Szalay (2011) afirma isso ao dizer que "a ciência moderna está se aproximando do ponto onde a história dos algoritmos e ferramentas computacionais combinados com o pensamento computacional se tornarão tão indispensáveis como a matemática". Mais do que isso, a Computação Científica tem mudado o modo de se fazer ciência e já é amplamente reconhecida como seu terceiro pilar, junto da teoria e da experiência. Isso é notável, principalmente, em algumas situações em que a infraestrutura computacional se torna o próprio laboratório de pesquisa para os cientistas (SZALAY, 2011; SAGER; MOMBAUR; FUNKE, 2013; KNIGHT; POULSON, 2013).

A metodologia científica tradicionalmente se baseia nos pilares da teoria e da experiência. Mormente nas ciências naturais, os teóricos se ocupam por descrever os fenômenos estudados com uma linguagem matemática, abstraindo assim o seu aspecto quantitativo que será objeto de análise; e os experimentalistas, por sua vez, realizam os experimentos necessários para validação das teorias em questão. A Computação Científica[2], ou *e-Science* como é chamada em alguns contextos, se apresenta como uma metodologia que está na intersecção da pesquisa científica, engenharia e Computação e apoia esses dois pilares. Quando os modelos teóricos se tornam demasiados complexos para uma solução convencional, ou quando as experiências se tornam muito perigosas ou proibitivas, o uso de modelos e simulações computacionais aparecem como a única opção viável para investigação e validação desses estudos. Esse caminho metodológico já trouxe avanços extraordinários sob vários aspectos e escalas do mundo físico, em áreas que vão desde a genética à física nuclear. Nesse sentido, também é importante destacar que embora inicialmente a Computação Científica estivesse focada em problemas numéricos, atualmente os pesquisadores usam as técnicas computacionais para um contexto muito mais amplo (KNIGHT; POULSON, 2013; PEREZ; GRANGER; HUNTER, 2011; SZALAY, 2011).

O desenvolvimento de um modelo computacional que implemente um modelo teórico é um trabalho interdisciplinar e que, geralmente, envolve profissionais de várias áreas. De forma geral, há três grandes funções a serem desempenhadas para construção desse tipo de modelo. Primeiramente, é necessário um especialista na área de pesquisa em questão, que trabalhe na construção do modelo teórico. Em um segundo momento, alguém que realize o trabalho de matemática aplicada, para formulação matemática do modelo. A terceira especialidade exigida é a de cientistas da Computação que estejam habituados com os algoritmos e máquinas que melhor se adaptam a solução dessas formulações teóricas e matemáticas. Todavia, é comum que todos que trabalham com Computação Científica estudem os modelos, algoritmos, *software* e *hardware* que juntos são utilizados no processamento computacional e na análise dos resultados (KNIGHT; POULSON, 2013).

[2] Daqui em diante será usado majoritariamente o termo Computação Científica para se referir a essa realidade.

Como mencionado, a disponibilidade de grandes conjuntos de dados e a capacidade de analisá-los computacionalmente estão mudando significativamente algumas metodologias científicas. Recentemente, esse processo está conhecendo um novo salto de ordem de grandeza em termos de armazenamento e de processamento, comumente chamado de *Big Data*. Szalay (2011) afirma que "nós estamos vivendo um tempo em que uma interessante transição está acontecendo na ciência, similar a revolução dos *Beowulf* na década de 90". Esse novo paradigma faz com que surjam novas soluções computacionais que tentam corresponder a essas novas necessidades. Tradicionalmente, as soluções em Computação Científica vinham se polarizando em dois quesitos, o desempenho e a usabilidade. Diante desse novo cenário, onde quantidades maciças e heterogêneas de dados estão disponíveis para análises sob a visão não apenas de sua ciência de origem, mas de várias outras abordagens científicas, as ferramentas e técnicas computacionais precisam ser flexíveis o bastante para oferecer mais do que apenas desempenho e usabilidade. É sobre esse cenário que Seidel (2012) afirma que:

> códigos científicos complexos e os conjuntos de dados que eles geram estão na necessidade de um ambiente de categorização sofisticado que permita a comunidade armazenar, pesquisar, e melhorar metadados em um sistema dinâmico e aberto. (SEIDEL, 2012)

É nesse contexto que se vê a crescente adoção de tecnologias que não estão voltadas somente para um desempenho segundo as demandas exigidas e uma usabilidade que permita a interação com os usuários finais das ferramentas. Elas, agora, valorizam a tão necessária interoperabilidade de informações, condição fundamental para que haja a construção de conhecimento. Nesse sentido, o uso de uma infraestrutura dinâmica e escalável, como a oferecida pelas plataformas da Computação em Nuvem, aos moldes de uma *Infrastructure as a Service* (IaaS), se mostra muito promissora sob a perspectiva do método científico. Isso é notável no que tange a reprodutibilidade de ambientes inteiros de pesquisa, que podem ser copiados, reproduzidos e melhorados, conforme se exija as necessidades da investigação (PRODAN; SPERK, 2013).

Interessante também é a adoção de um *framework* baseado em padrões compartilhados de interoperabilidade, responsável por fornecer um ambiente que lide com a heterogeneidade de dados, metadados e suas fontes. É importante lembrar, como aponta Szalay (2011), que as "análises dos dados de outras disciplinas (tal como as ciências ambientais) devem abranger milhares de dados distintos, com formatos incompatíveis e metadados inconsistentes". Para traçar um pouco melhor esse quadro de conceitos, tecnologias e ferramentas usadas para solução dos problemas relacionados à Computação Científica, esta seção explora algumas dessas iniciativas promissoras, como é exposto nas próximas subseções.

2.1.1 O uso da Computação em Nuvem

A Computação em Nuvem como é conhecida hoje surge dentro do contexto de desenvolvimento, e acúmulo, de *know-how* proporcionado por diversas tecnologias ligadas à Computação Paralela e Distribuída. Talvez por seu desenvolvimento aparecer de forma tão paulatina, uma linha que delineasse bem o seu escopo e uma definição de consenso tenha sido tão difícil de alcançar no início. Foster et al. (2008) destacam que o conceito de Computação em Nuvem não é algo propriamente novo e sim uma intrincada associação de tecnologias. Reforçam que foram necessários mais de trinta anos de desenvolvimento de Computação em *Grid*, em *Clusters*, e de sistemas distribuídos em geral, para se chegar ao estado da arte necessário para o surgimento desse paradigma.

Uma definição é proposta por Buyya et al. (2009) ao afirmarem:

> A nuvem é um tipo de sistema paralelo e distribuído que consiste em uma coleção de computadores interconectados e virtualizados, que são fornecidos dinamicamente e apresentados como um ou mais recursos unificados de computação, com base nos acordos de nível de serviço estabelecidos através da negociação entre o provedor de serviços e os consumidores. (BUYYA et al., 2009)

É interessante notar que essa definição busca destacar dois aspectos. Por um lado, tem-se a descrição dos recursos de Computação e da virtualização assumida por eles, e por outro, apresenta-se a forma de se disponibilizar essa infraestrutura computacional como um serviço. É principalmente em torno desses aspectos que ocorre a discussão a respeito da definição de Computação em Nuvem. Buyya, Pandey e Vecchiola (2009) deixam isso mais claro ao dizerem que "Computação em Nuvem entrega infraestrutura, plataforma e *software* (aplicações) como serviços [...]". Fazem nota também de que se consolidou na indústria os termos *Infrastructure as a Service* (IaaS), *Platform as a Service* (PaaS) e *Software as a Service* (SaaS) para descreverem essa modalidade de fornecimento de recursos de computação. É importante ressaltar aqui que o termo Computação em Nuvem se refere aos recursos e aplicações entregues sobre a Internet e aos *data centers* que fornecem e dão suporte a esses serviços (BUYYA; PANDEY; VECCHIOLA, 2009; ARMBRUST et al., 2010).

Em termos administrativos, quando uma organização afirma estar operando a partir da Nuvem quer dizer que está terceirizando as suas demandas computacionais para outra parte, que gerencia, mantém e fornece esses recursos como uma prestadora de serviços (BUSHHOUSEN, 2011). Isso traz inúmeras vantagem como relatam Armbrust et al. (2010) ao explicarem que os desenvolvedores e pesquisadores diante de uma iniciativa potencialmente inovadora não precisam mais se preocupar se esse serviço será subestimado ou superestimado. Pois, diante de uma situação ou outra, os recursos de infraestrutura de TI, que são alocados virtualmente, podem ser facilmente redimensionados devido à

escalabilidade da Computação em Nuvem. Dessa forma, os interessados nessas iniciativas, como ressaltam os autores, "já não necessitam de grandes investimentos de capital em *hardware* para implantar seu serviço ou em gasto humano para operá-lo" (ARMBRUST et al., 2010).

Legalmente, *Infrastructure as a Service* (IaaS) significa que esse terceirizado executa os processos de armazenamento, gerenciamento, processamento e transferência de dados em seu *hardware*, porém, o cliente continua detentor da propriedade e dos direitos sobre as aplicações e os dados. De modo que o cliente apenas não mais provê diretamente o *hardware* necessário para manter a disponibilidade desses recursos computacionais. *Software as a Service* (SaaS) é a modalidade de acordo em que se fornece aos clientes a capacidade de uso das aplicações que estão hospedadas na infraestrutura da Nuvem. Os clientes não têm acesso aos detalhes de implementação e infraestrutura, e estão normalmente limitados às simples configurações de usuários. Por sua vez, o termo *Platform as a Service* (PaaS) é associado àqueles acordos de serviços que preveem que seja fornecido ao cliente a capacidade de realizar a implantação dos *software* e aplicações adquiridos ou desenvolvidos por ele próprio. Nesse caso, o cliente continua não tendo acesso aos detalhes de infraestrutura, porém, lhe é concedido acesso as aplicações implantadas por ele e às configurações do ambiente de hospedagem (BUSHHOUSEN, 2011; MELL; GRANCE, 2011).

Esse apanhado teórico sobre a organização da Computação em Nuvem foi realizado para uma maior compreensão do seguinte ponto. Na comunidade científica, a última década foi marcada por um cenário característico. Por um lado, as pesquisas de maior porte possuíam recursos para alocar e manter uma infraestrutura de supercomputadores para o tratamento dos dados e processamento dos modelos computacionais. Em contrapartida, a maioria das pesquisas de menor porte não podendo manter esse tipo de ambiente foram obrigadas a adotar *clusters* comerciais e *grids*. Nesse contexto, a Computação em Nuvem surge como uma modalidade de fornecimento de recursos computacionais que se encaixa de forma proporcional na demanda dos projetos de pesquisa, e apenas durante o período de tempo necessário. Ostermann et al. (2010) afirmam que "O paradigma da Computação em Nuvem reserva uma boa promessa para a fome de performance da comunidade científica". Contudo, permanecem realistas ao notarem que, por ter sido inicialmente desenvolvida mais para área comercial, a Nuvem ainda se mostra insuficiente para uma demanda em larga escala e de alto desempenho como é característico do processamento de Computação Científica. Entretanto, ela se apresenta como adequada para pesquisas que precisem de recursos de computação de forma imediata e por um período determinado de tempo. Fox (2011) ressalta que, embora as Nuvens tenham de início adotado um *hardware* segundo a demanda das Aplicações Web, os primeiros clientes delas, e tenham mostrado resultados fracos, atualmente um "Novo *hardware* está configurado para um melhor desempenho das aplicações científicas".

Outro ponto a se notar é que os serviços que rodam na Nuvem estão se unindo às infraestruturas convencionais de Computação e juntas buscam oferecer plataformas viáveis para pesquisa científica e ensino. De modo que se torna "fundamental entender as formulações e modos de uso que sejam significativos em tal infraestrutura hibrida, juntamente com conceitos fundamentais, desafios tecnológicos e meios pelos quais essas aplicações possam efetivamente utilizar as Nuvens" (PARASHAR; THIRUVATHUKAL, 2013).

Atualmente, considerando vantagens e limitações da Computação em Nuvem usada na Computação Científica, é possível destacar como umas das principais vantagens aquilo que o Tim Bell, diretor do centro de infraestrutura e serviços operacionais do LHC, resume ao relatar que "No passado, quando eles pediam por *hardware* físico, eles esperavam por semanas, agora eles podem pedir por uma máquina virtual e conseguir algo no tempo que leva para pegar um copo de café" (DRAKE, 2014).

2.1.2 O uso de Infraestruturas Virtuais

Com advento da Computação Científica e com o uso da Computação em Nuvem ficou claro que as demandas dos pesquisadores se modificaram justamente acompanhando as novas ferramentas e metodologias de se fazer Ciência. Diante dessa perspectiva é proveitoso apontar alguns trabalhos que tratam do uso de infraestruturas virtuais adaptadas à Nuvem e ao uso de Computação Científica nos projetos de pesquisa. Seguem abaixo, algumas de suas iniciativas e conceitos mais importantes.

2.1.2.1 Iniciativas

Hey e Trefethen (2005) destacam que os cientistas na época já buscavam meios de colaboração para além do convencional cenário da Web. Eles ressaltam isso ao dizerem que os pesquisadores para além de serem capazes "de acessar informações de diferentes lugares, eles querem ser capazes de integrar, associar e analisar informações de várias fontes de dados diferentes e distribuídos [...] e acessar e controlar recursos computacionais e equipamento experimental de forma remota." (HEY; TREFETHEN, 2005). Nesse cenário, cada vez mais automatizado, os metadados surgem como meio importante para eles poderem utilizar ferramentas de minerações de dados que lhes permitem entender e analisar melhor as grandes massas de informações. Elas já eram estimadas na ordem de petabytes de dados científicos, provenientes das mais diversas fontes de instrumentação científica, como experimentos de alto desempenho, simulações de supercomputadores, redes de sensores, levantamentos remotos via satélite, entre outros. É importante notar que essa grande tendência de colaboração científica, que por sua natureza é distribuída e que, por ser multidisciplinar, se realiza em diversas áreas do conhecimento, pode se beneficiar de uma infraestrutura em comum. Para tentar contribuir com uma solução

para essa demanda crescente, um projeto descrito por Hey e Trefethen (2005) tinha por objetivo reunir em uma Organização Virtual, do inglês *Virtual Organization* (VO), dois grupos de pesquisas de duas universidades diferentes. Nesse ambiente, eles poderiam compartilhar dados coletados e modelos desenvolvidos por ambas as instituições, além de compartilharem os recursos computacionais da infraestrutura subjacente a todo esse ambiente. Esse conceito de VO é importante para esses tipos de ambientes virtuais, como será mostrado em detalhes mais à frente. Essencialmente, essa inciativa se baseia no fornecimento de uma infraestrutura virtual de TI por demanda, que comumente é designada como *e-Infrastructure* ou *cyberinfrastructure* (nos Estados Unidos). Essas infraestruturas virtuais se comportam como um *pool* de serviços que são prestados para suprir as necessidades de um ambiente de pesquisa científica (HEY; TREFETHEN, 2005). É sobre as linhas principais desse assunto que esta seção busca discorrer, comentando os conceitos-chaves ao mesmo tempo em que busca trazer alguns trabalhos de relevo.

Uma iniciativa que se pode destacar é a de Han et al. (2010) que propõem uma outra *e-Infrastructure* baseada em um *Grid* de Alto Desempenho, que busca resolver os problemas relacionados a chamadas de emergência em larga escala, inicialmente focadas em incêndios. O projeto pode ser descrito sucintamente como uma arquitetura de sistema em que os dados são coletados por sensores em tempo real, filtrados e armazenados; nesse ambiente também ocorre o desenvolvimento de modelos, que são colocados em funcionamento em uma infraestrutura de alto desempenho computacional.

Um projeto de envergadura transnacional é comentado por Balaž et al. (2011), eles relatam a natureza do *Grid* regional *South-East European GRid e-Infrastructure Development* (SEE-GRID), que é mantido pela Comissão Europeia. O foco de trabalho dessa poderosa infraestrutura se centrou nas comunidades de sismologia, meteorologia e de ciências ambientais. Os autores não deixam de discutir sobre a mudança nos paradigmas metodológicos de se fazer Ciência, destacando que a transição da Ciência tradicional para a *e-Science* é impulsionada pela demanda de uma resposta cada vez mais forte em termos de recursos computacionais. É preciso ter em mente que para descreverem melhor os problemas do mundo real, as pesquisas estão se utilizando das melhorias na precisão de observação dos aspectos quantitativos de seus objetos de estudos. Isso acontece seja por meio de sensores mais sofisticados e detalhistas ou de simulações numéricas mais robustas. O SEE-GRID, e as iniciativas de *e-Infrastructure* de forma geral, visa suprir o gargalo existente entre essa demanda e os recursos e técnicas computacionais atualmente existentes.

Outro projeto de destaque é Low Frequency Array (LOFAR) que disponibiliza uma rede radiotelescópios de baixa frequência e sua respectiva *e-Infrastructure* no norte da Europa. Essa inciativa busca prover uma infraestrutura que cuide do gerenciamento, analise e armazenamento de dados científicos produzidos por eles. Para isso, cada local

envolvido no projeto fornece uma capacidade de armazenamento e opcionalmente outra de processamento (HOLTIES; RENTING; GRANGE, 2012).

Um outro trabalho, desenvolvido por Eriksson e Goldkuhl (2013), trata do desenvolvimento de um *e-Infrastructure* para governo eletrônico. Eles esclarecem que apesar das infraestruturas cibernéticas entregarem os serviços de TI com interoperabilidade o conhecimento de como desenvolver uma infraestrutura virtual pública ainda é limitado. Nesse trabalho, ao tratarem do desenvolvimento do estágio inicial de um projeto de *e-Infrastructure* para o setor público, eles listam seis aspectos no desenvolvimento de infraestruturas cibernéticas que são interessantes destacar: os aspectos legal, econômico, organizacional, técnico, informacional e contratual (ERIKSSON; GOLDKUHL, 2013).

Agregando ainda mais às iniciativas desse cenário, Lecca et al. (2011) afirmam que as *e-Infrastructures* prometem modificar fortemente o modo como os sistemas sensoriais funcionam, juntamente com a maneira de coletar e persistir dados. Além disso, também o jeito como os modelos formados a partir desses dados serão simulados e visualizados. Eles comentam isso, pois a iniciativa deles atua dentro de um programa de infraestruturas virtuais promovido pela União Europeia. Esse programa já busca aglutinar vários esforços para promover essas infraestruturas de conhecimento, mirando em pesquisa, educação e inovação para as próximas décadas. Esses autores também destacam nesse artigo que os *grids*, em um nível inferior de abstração de tecnologias, são componentes fundamentais para construção das *e-Infrastructures*. Nesse trabalho, eles se mostraram bem sucedidos na predição de inundações, manejo de recursos terra-água e no levantamento hidrológico do Mar Morto. É colocado em evidência que as demandas nessa área exigem não apenas o acesso direto a banco de dados geograficamente distribuídos e estruturalmente heterogêneos, mas também aos recursos computacionais para gerenciar e utilizar essas informações. Lembrando que esses experimentos de sucesso obedeceram às mais rígidas normas da sociedade civil no que concerne aos riscos naturais e antrópicos.

Levando em consideração os recursos computacionais, destaca-se que os requisitos de *hardware* crescem cada vez mais, em termos de processamento e armazenamento. Porém, apesar desse crescimento, o gargalo existente há décadas entre demanda e oferta de recursos continua existindo. Isso limita a capacidade dos pesquisadores de tomar decisões para identificar, reunir e analisar os dados relevantes de uma área de interesse. Nesse contexto, um ponto positivo das infraestruturas virtuais está nas interfaces amigáveis para portais Web e outras comunidades virtuais, que foram desenvolvidos para responder as necessidades da comunidade-alvo de pesquisadores (LECCA et al., 2011)(COSSU et al., 2010 apud LECCA et al., 2011).

2.1.2.2 Conceitos

Todas as iniciativas que foram citadas estão dentro de um bojo de conceitos, que as geram ou surgem delas. Por isso, alguns conceitos fundamentais ligados à elas são explorados nos parágrafos a seguir.

Walker et al. (2011) destacam que a literatura científica precedente sobre o assunto da Computação Científica costuma classificá-la e de certa forma simbolizá-la, junto com todos seus componentes – caso das infraestruturas virtuais – como um quarto paradigma do método científico, expressão de muita importância para Jim Gray. Isso é visto em Hey, Tansley e Tolle (2009), em um artigo de um livro editado por eles mesmos e dedicado exclusivamente a esse assunto. Os autores fazem certas distinções no desenvolvimento do método científico. Eles afirmam que em um primeiro momento a Ciência se concentrou em seu componente empírico, esse seria o primeiro paradigma. O segundo paradigma teria acontecido quando o aspecto teórico do método científico ficou em maior evidencia há alguns séculos atrás. Depois disso, o terceiro paradigma teria surgido devido à dificuldade de uma solução analítica para o modelos teóricos, quando os cientistas apelaram para soluções numéricas por meio da simulação computacional no século XX. Já o quarto paradigma seria a consolidação, no início do século XXI, do uso de tecnologias computacionais em todo o processo de desenvolvimento científico, desde a captura dos dados até a sua análise final. Enfatiza-se que os *software* precisam viabilizar o quarto paradigma de descobertas promovido pela Computação Científica, de modo a incluir ferramentas de planejamento, agendamento e monitoramento de dados em aplicações de larga escala e distribuídas.

Nesse sentido, Cheptsov et al. (2012) afirmam que nos últimos anos as redes e a Computação de Alto Desempenho já começam a tornar possível essas infraestruturas em larga escala. Destacam que certas fundações já entregam ambientes de Computação Distribuída com milhares de núcleos de processamento e petabytes em armazenamento para essas aplicações de alto desempenho. Apesar desse grande crescimento da capacidade de computação e armazenamento das soluções de TI, ainda continua grande o desencontro entre essas infraestruturas virtuais e as aplicações de Computação Científica que a utilizam. Elas ainda são mais usadas de forma direta com instrumentos, sensores e equipamentos de laboratórios. O uso remoto desses instrumentos pode melhorar pela integração delas com as infraestruturas já existentes. De modo a alcançar o que os autores almejam, uma *e-Infrastructure* que crie um espectro absolutamente novo de oportunidades para um certo número de projetos de pesquisa (CHEPTSOV et al., 2012).

Mais recentemente, em relação à transição dos métodos tradicionais para a Computação Científica, Barjak et al. (2013) expõem o crescimento constante desse esforço e os significativos fundos que os projetos ligados à *e-Infrastructures* costumam receber por parte de órgãos de fomento à pesquisa de diversos países. Para se ter uma ideia, o orçamento anual desses programas, no Reino Unido, foi estimado em £ 170 milhões.

Também ressaltaram a abrangência dessa iniciativa que alcança domínios de pesquisa que se estendem desde Ciência da Computação, física nuclear e outras ciências exatas passando por biomedicina, climatologia até ciências sociais e outras humanidades. Eles elencam como característica comum desses projetos de *e-Infrastructure* – característica que é possível tomar como definição para Computação Científica – o fato de que eles

> recorrem à recursos digitais distribuídos geograficamente – tais como dados, poder de computação, tecnologia de visualização e armazenamento –, a fim de fornecer serviços que permitam o compartilhamento de recursos e ferramentas colaborativas essenciais para pesquisa em ambientes distribuídos. (BARJAK et al., 2013)

Nesse mesmo trabalho, os autores destacam uma figura importante de governança de TI que está relacionada a melhor gestão de *e-Infrastructures*. Essa figura é geralmente designada pelo termo VO que expressa uma abstração de uma entidade dentro das *e-Infrastructures*. Esse termo pode ser definido como sendo "um grupo de indivíduos cujos membros e recursos podem estar dispersos geograficamente e institucionalmente, ainda que funcione como uma unidade coesa pelo uso de uma infraestrutura cibernética" (BARJAK et al., 2013). Essa figura é muito importante como elemento organizador dentro da hierarquia de componentes adotados por iniciativas como o *gCube*, que é utilizado neste trabalho.

O trabalho de Assante, Candela e Pagano (2013), por sua vez, põe em evidência as mencionadas mudanças que os suportes para se fazer Ciência estão sofrendo. A multidisciplinaridade, a intercomunicação em rede e a adoção de novos métodos com o advento da Computação Científica resulta em envolvimento de várias áreas do saber nas pesquisas, de forma a compor um complexo cenário. A comunicação tradicional do meio acadêmico já não é suficiente para que os pares consigam reproduzir os resultados alcançados pelos trabalhos dessa natureza. Faz-se necessário o acesso aos diversos artefatos produzidos e descobertos durante o fluxo do trabalho científico, como, por exemplo, conjunto de dados, ferramentas de análise e outros métodos. Nesse cenário é que uma infraestrutura virtual comum consegue ajudar efetivamente os pesquisadores envolvidos nessas atividades. Para expor a viabilidade dessa comunicação entre pesquisas, esses autores lançam mão de um outro conceito que possui grande importância no gerenciamento de infraestruturas virtuais. Trata-se dos Objetos de Pesquisa (do inglês, *Research Objects*) que podem ser definidos como "uma abstração para a comunicação, compartilhamento e reutilização de resultados de pesquisa" (ASSANTE; CANDELA; PAGANO, 2013). Eles consistem em um agregado de várias partes de arquivos digitais de diversos tipos, que vão desde arquivos binários, passando por séries temporais, mapas georreferenciados até outros objetos mais complexos. Além disso, um ambiente de produção de Objetos de Pesquisa foi apresentado nessa iniciativa, esse ambiente inclui uma área de trabalho para troca de itens, uma área de edição de Objetos de Pesquisa e um motor de *workflow* para gerenciar o desenvolvimento desses itens (ASSANTE; CANDELA; PAGANO, 2013).

Candela et al. (2013a) afirmam que dentre as soluções tecnológicas existentes para lidar com a *e-Science* as *e-Infrastructures* desempenham o papel principal, em especial o desempenhado pelas *Hybrid Data Infrastructures* (HDI). Essa última categoria de infraestruturas virtuais é descrita por Candela, Castelli e Pagano (2012) como "uma nova solução, mais eficaz para a gestão dos novos tipos de conjunto de dados científicos. Ela assume que várias tecnologias, incluindo *Grid*, nuvens privadas e públicas, podem ser integradas". Seu objetivo, como o das outras infraestruturas cibernéticas, é fornecer recursos como serviços, por meio dos conceitos de aplicações como serviço ('SaaS'), *hardware* e plataforma como serviço ('IaaS' e 'PaaS'). Ademais, uma HDI oferece uma grande quantidade de serviços de dados e gerenciamento de dados, além de poder utilizar outras infraestruturas. Nesse contexto, é utilizado a figura dos Ambientes Virtuais de Pesquisa, do inglês *Virtual Research Environments* (VRE), que são ambientes feitos para Web e desenvolvidos para reunir todas as ferramentas necessárias para a produção de uma investigação científica (CANDELA et al., 2013a).

2.1.2.3 gCube

Uma tecnologia que implementa os conceitos já citados (subseção 2.1.2.2) sobre infraestruturas virtuais é o *gCube*, que é uma importante ferramenta para viabilização dessas ações. Ele se apresenta como uma forma de viabilizar a construção e operação de uma HDI. Pode ser descrito sucintamente como um sistema de *software* desenvolvido com vários componentes responsáveis por prover o gerenciamento, armazenamento e processamento de dados, dentre outras funcionalidades, como serviços. Para maior compreensão de sua natureza seus idealizadores apontam três perspectivas sob as quais o *gCube* pode ser entendido.

A primeira perspectiva é a de *um provedor de uma infraestrutura de dados*. Explicado de outra forma, pode-se dizer que é uma maneira de transformar as tecnologias e infraestruturas em serviços ou ferramentas. O problema para se entregar uma infraestrutura virtual é que isso exige a integração de diversas tecnologias. Para possibilitar isso se deve enfrentar a multiplicidade de fornecedores de aplicações complexas, os diversos paradigmas em que essas tecnologias foram desenvolvidas, as tecnologias de *middleware* para abstração da heterogeneidade do *hardware* e os vários padrões existentes para uma mesma camada de *software*. O *gCube* supera essas dificuldades pela adoção de soluções que abstraem as diferenças de tecnologias, padrões e protocolos e por um sistema de suporte a erros, que abrange uma grande quantidade de potenciais problemas. É interessante ressaltar que embora abstraia essa diversidade de tecnologias o *gCube* não as esconde, não atua como outra camada do meio. Porém, oferece uma visão comum para acesso, desenvolvimento, monitoramento e apresentação dessas tecnologias (GCUBE CONSORTIUM, 2015).

Uma segunda forma de ver o *gCube* é a sob a perspectiva de *um framework para*

dados e processamento. O *gCube* aqui se apresenta como um *framework* projetado para abstrair, como já destacado, uma grande diversidade de tecnologias responsáveis pelo armazenamento, processamento e gerenciamento de recursos. Isso tudo sobre uma Nuvem ou *Grid* que uma *middleware* provisiona. Essa abstração apresenta essas tecnologias como um grupo organizado e homogêneo de API's e serviços. Essa arquitetura consegue entregar no todo: (a) acesso ao armazenamento de informações em diversos formatos, dependendo de seus propósitos, como pacotes de *software*, grandes conjuntos de dados científicos em tabelas, series temporais com OLAP para manipulá-las, objetos de documentos estruturados e árvores de hierarquia desses objetos, dados georreferenciados e arquivos de texto simples; (b) um motor de execução de processamento chamado PE2ng, que possibilita o processamento de dados guiado por modelos de *workflows*; (c) um motor de transformações para enfrentar o problema relacionado a transformação dos dados que muitas vezes estão sob vários formatos, ele está posto sobre o PE2ng e pode ser adicionado a ele como um componente; (d) e, finalmente, um gerenciador de VREs, pelos quais grupos de usuários podem de forma controlada e organizada fazer a integração dos tão necessários artefatos da Computação Científica, como dados, serviços para os mais diversos fins, recursos computacionais e diversas outras ferramentas. Sobre os VREs é importante destacar que eles possibilitam a cooperação entre pessoas em tarefas. Essas tarefas podem ser o enriquecimento dos dados com alguns formatos de metadados como, por exemplo o uso de um vocabulário controlado ou esquemas de validação. As técnicas que permitem adição de metadados são o fundamento para mitigação do problema de heterogeneidade semântica. Além disso, as tarefas compartilhadas se estendem à avaliação dos dados, bem como os produtos gerados pelo processamento desses dados ou simulações. Em suma, permitem a integração, compartilhamento e cooperação no armazenamento e uso dos dados (e seus metadados que os fazem semanticamente viáveis para outros grupos); também possibilitam os processamentos aplicados sobre eles pelos *software*, que igualmente podem ser compartilhados (GCUBE CONSORTIUM, 2015).

A última perspectiva ressaltada é a de *aplicação científica*, que se resume como "um modo de se entregar aplicações científicas na nuvem." (GCUBE CONSORTIUM, 2015). Ela se torna mais clara quando se tem em perspectiva as várias aplicações da área de piscicultura que já foram implementadas usando o *gCube*. Agora essas aplicações já fazem parte das atuais revisões do *gCube*. Porém, além dessas, outras aplicações científicas podem ser destacadas. Foi desenvolvido um *software* que atua como um ambiente de acesso e organização de conjuntos de informações, que são descritas como objetos. Dentre os quais é possível ver objetos de informação, séries temporais, consultas, arquivos, modelos etc. Em termos de uso, ele se assemelha à um gerenciador de arquivos. Existe, também, um *framework* para séries temporais no qual se disponibiliza várias ferramentas para se analisar e se operar com estatísticas multidimensionais esses dados. Além de um suíte para modelagem de nicho ecológico, que permite predições da distribuição de espécies marinhas

com base em dados sobre as espécies e as condições ambientais (GCUBE CONSORTIUM, 2015).

Com relação aos serviços oferecidos por essa implementação de infraestrutura virtual, pode-se destacar alguns de maior relevância. Um importante serviço do *gCube* é o *Live Research Objects Environment* (Ambiente de Objetos de Pesquisa Direta, em tradução livre), que é um dos serviços responsáveis pelo gerenciamento dos dados. Ele implementa alguns conceitos já citados, é um espaço de trabalho virtual, para troca desses objetos de informação, um ambiente de edição de *Live Research Objects* e um gerenciador de fluxo de trabalho de *Research Objects*, que relaciona um determinado fluxo de trabalho com um *Research Object*. É necessário lembrar que a ideia geral do *gCube* é que ele age como uma abstração para diversas tecnologias de armazenamento, processamento e gestão de recursos, de modo a apresentar para o usuário um conjunto homogêneo de aplicações e serviços. Para isso ele oferece opções de abstração sobre essas diferenças, escalonando apenas os recursos disponibilizados em alto nível, também contando com certa tolerância a falhas (ASSANTE; CANDELA; PAGANO, 2013).

Como já foi mencionado, uma forte figura que surge na gestão desse tipo de infraestrutura é a do Ambiente de Pesquisa Virtual. Alguns detalhes podem ser acrescentados a seu respeito para melhor compreensão. São ambientes desenvolvidos para Web, orientados para comunidade, flexíveis e seguros, modelados especialmente para o apoio da Computação Científica. Uma HDI oferece um serviço de gerenciamento desses VREs que por sua própria natureza abarca a implantação, o monitoramento e a operacionalização de VREs. É importante destacar, que é em uma VRE que os diversos pesquisadores e colaboradores podem trocar seus recursos de dados, processamento, e demais ferramentas científicas utilizadas na produção de sua pesquisa. Os recursos compartilhados estão nos dois aspectos da computação, no *hardware* e no *software*, incluindo também outras infraestruturas virtuais, que estejam ou não na Nuvem. Pode-se dizer também que uma HDI tem por objetivo promover esse tipo de comunidade com serviços de gerenciamento de dados por meio desses VREs. Nesse contexto, um Ambiente de Pesquisa Virtual atua de modo viabilizar essa integração e interoperacionalidade de diversas ferramentas, por meio de uma maior clareza semântica. Isso é obtido pela possibilidade de uso de mapeamento de *schemas*, tesauros, ontologias e demais tecnologias para lidar com metadados e estrutura da informação. Além disso, um VRE permite o processamento e análise dos dados lá disponíveis, que por sua vez podem ser compartilhados junto de suas ferramentas e processos utilizados, como já mencionado (CANDELA et al., 2013a).

Nas palavras de Candela (2013b) um VRE é algo abrangente, que pode ser melhor definido como um

> sistema com as seguintes características distintivas: (i) é um ambiente de trabalho feito para Web; (ii) que é feita sob medida para atender

> as necessidades de uma comunidade de prática (Lave e Wenger, 1991); (iii) espera-se prover uma comunidade de prática com toda a gama de produtos necessários para realizar os objetivos da comunidade; (iv) é aberta e flexível no que diz respeito à oferta de serviços em geral e tempo de vida; e (v) promove bem precisa e controlada partilha de ambos os resultados da pesquisa, intermediários e finais, garantindo a propriedade, proveniência e atribuição. (CANDELA, 2013b)

Sobre esse importante conceito vale destacar que se tratam de ambientes arrojados e adaptados a essa nova forma de se fazer Ciência baseada no paradigma que abarca a Computação Científica, cada vez mais comuns nos dias atuais. Os VREs permitem essa grande dinamicidade no compartilhamento de recursos, dados, ferramentas e técnicas, de modo a transcender as fronteiras entre instituições de ensino, dentre outras organizações (CANDELA, 2013b).

2.1.2.4 D4Science e EUBrazilOpenBio

Ligado ao projeto *D4Science*, fundado em colaboração com órgãos da União Europeia e que foi responsável pelo desenvolvimento do *framework gCube*, surgiu uma organização de mesmo nome. Essa organização oferece uma infraestrutura para dados híbridos, do inglês *Hybrid Data Infrastructures*, baseado na tecnologia *gCube* para diversos colaboradores do âmbito científico. Essa infraestrutura foi desenvolvida para prover a colaboração entre projetos de pesquisa em vários âmbitos. Ela apresenta como meta mais importante a simplificação do provisionamento de uma infraestrutura virtual agregadora de infraestruturas que possam viabilizar o provisionamento de VREs.

A partir do início com o projeto homônimo, essa iniciativa se expandiu em parceria com outros projetos como o *D4Science-II*[3] (2009-2011), ENVRI[4] (2011-2014), EUBrazilOpenBio[5] (2011-2013) e iMarine[6] (2011-2014). Atualmente os recursos de *hardware* e a infraestrutura estão em três locais principais, a saber: em Pisa, em um laboratório do Conselho Nacional de Pesquisa da Itália[7], em Roma, no Grupo Engineering[8] e em Atenas, na Universidade de Atenas[9] e na Communication and Information Technologies Experts S.A.[10] (D4SCIENCE INFRASTRUCTURE, 2015a).

Como uma HDI, o *D4Science* oferece uma gama de mais de 500 componentes de *software* e integra virtualmente mais de 50 provedores de dados, além de todas as funcionalidade e serviços disponíveis no *gCube*. Isso tudo dentro de um ambiente coerente fornecido pela infraestrutura virtual (D4SCIENCE INFRASTRUCTURE, 2015b).

[3] <http://www.d4science.eu/>
[4] <http://envri.eu/>
[5] <http://www.eubrazilopenbio.eu/>
[6] <http://www.i-marine.eu/>
[7] <http://nemis.isti.cnr.it/>
[8] <http://www.eng.it/>
[9] <http://www.madgik.di.uoa.gr/>
[10] <http://www.cite.gr/>

Para explorar os recursos oferecidos pela infraestrutura virtual *D4Science* é necessário ingressar com um pedido junto aos mantenedores dela. Existem três modalidades para se usar a HDI do *D4Science*. A primeira é sob o modelo de "*As-a-User*", que é indicado se a pessoa possui a necessidade de armazenar, acessar, e se achar oportuno compartilhar arquivos e conjuntos de dados de seus trabalhos. Além de ter acesso ao trabalho, ferramentas e dados de muitos outros usuários das diversas organizações e VREs. A segunda é sob o modelo de "*As-a-Group*", que propícia a criação de um ambiente de pesquisa (VRE) dedicado para os fins de um grupo de pesquisadores. Nesse ambiente se pode implantar e ter acessos as diversas aplicações criadas no âmbito dessas pesquisas, além dos conjuntos de dados organizados pelos colaborados do grupo. Há nesse cenário todas as ferramentas necessárias para a gestão e coordenação do acesso e papéis das pessoas do grupo. E a terceira forma é como uma comunidade sob o modelo "*As-a-Community*". Nesse caso, é indicado para a integração e o gerenciamento de uma comunidade de maior porte, onde haja a expectativa de atividade em um período razoável de tempo. Esse ambiente seria ideal para o compartilhamento de dados, recursos computacionais e de suas rotinas diárias de processamento. Também é possível, como no modelo anterior, a implantação de aplicações e ferramentas próprias (D4SCIENCE INFRASTRUCTURE, 2015c).

Uma iniciativa mais próxima, que se utiliza do conceito de HDI, da tecnologia *gCube* e do *D4Science*, é o *EUBrazilOpenBio*. Segundo Amaral et al. (2014) se trata de uma iniciativa para avançar "sobre as barreiras estratégicas em pesquisa da biodiversidade através da integração de dados de acesso livre e ferramentas de fácil utilização, amplamente disponíveis no Brasil e na Europa". Essa iniciativa é apresentada como um ambiente de Computação Científica integrado para cientistas da área de biodiversidades. "Em essência, ele [...] introduz a abrangência de recursos e capacidade infinita como principais características, com o objetivo de tornar os dados e serviços de gerenciamento de dados disponível *on-demand*." (AMARAL et al., 2014). Eles mantém o HDI *EU-Brazil*, que implementando os diversos conceitos já citados, como, por exemplo, o de *e-Infrastructure* e *Hybrid Data Infrastructure*, permite a entrega de recursos computacionais como serviços, de acordo com a demanda do grupo de pesquisa interessado.

De forma mais específica se observa a entrega de dados ambientais ligados a biodiversidade e o clima. Uma outra característica é que eles se utilizam de tecnologias e conceitos já desenvolvidos, tendo assim uma natureza agregadora. Como, por exemplo, na adoção do conceito de VRE e outras figuras de gestão já citadas. Sobre o VRE, os autores destacam que o ele possui uma pequena curva de aprendizagem. Amaral et al. (2014) citam outros projetos na Europa que estão em um ambiente semelhante ao deles, como o SCI-BUS, que trabalha na definição de um *framework* flexível para o desenvolvimento de *gateways* para uso científico; o *D4Science* que é a base para o *EUBrazilOpenBio*, e que utiliza a mesma tecnologia *gCube* e o conceito de VRE, dentre outros projetos. A fim de resumir algumas características da situação inicial de uma pessoa que passe a utilizar as

duas maiores iniciativas que foram citadas, gCube e D4Science, segue a Tabela 1.

Tabela 1: Tabela com características do gCube e do D4Science.

Características	**gCube**	**D4Science**
VO	Sim	Sim
VRE	Sim	Sim
Objetos de Pesquisa	Sim	Sim
Infraestrutura de hardware	Nuvem própria	Nuvem da organização
Provedores de dados	Próprios	Próprios e de outros colaboradores
Componentes de software	Próprios	Próprios e de outros colaboradores

Características, como as mostradas na Tabela 1, foram usadas como critérios para se escolher a infraestrutura virtual usada neste trabalho. Todavia, essa e outras escolhas estão detalhadas no Capítulo 3.

2.2 Heterogeneidade Semântica

A heterogeneidade semântica começa a se tornar um objeto de estudo mais frequente, no tocante ao armazenamento e gerenciamento de dados, quando nos anos 80 o número de bases de dados sofreu um crescimento vertiginoso. Passou-se a um cenário comumente descrito como "ilhas de informação". Isso fez com que as técnicas para comunicação entre os banco de dados da época se defrontassem com o problema de heterogeneidade dos vários tipos de bancos de dados existentes. Posteriormente, com a invenção e popularização da *World Wide Web* (Web), na década de 90, o que já era vasto passou para uma escala antes jamais imaginada, fazendo com que o problema adquirisse uma grande visibilidade e importância. Mas em que consiste a heterogeneidade semântica? E quais outras considerações podem ser feitas a respeito do tema, suas características e soluções? A heterogeneidade semântica, como será discutido na subseção 2.2.1, está relacionada a diversidade na interpretação e representação dos dados. Serão explorados também os tipos de diversidade semântica e como são caracterizados (subseção 2.2.2), os locais de ocorrência desse fenômeno (subseção 2.2.3) e algumas soluções adotadas para lidar com esse problema (subseção 2.2.4).

2.2.1 Definições

É observado que uma definição já é explorada no início dos anos 90, na comunidade de Ciência da Computação, como a conhecida formulação de Sheth e Larson (1990): "Heterogeneidade semântica ocorre quando há uma discordância sobre o significado, a interpretação ou a utilização a que se destinam os mesmos dados ou relacionados". Devido ao enfoque de seus trabalhos, vários pesquisadores começam a destacar um ou outro

aspecto dessa formulação sintética, desenvolvendo uma gama de classificações e buscando precisar melhor as causas desse fenômeno.

Ventrone e Heiler (1991) argumentam que uma das causas da heterogeneidade semântica se deve a "evolução do domínio". Isso ocorreria quando os significados dos objetos do mundo real – relacionados a dados armazenados digitalmente – se modificam, podendo fazer com que a massa de dados se torne um agregado de subdomínios semanticamente incompatíveis. De forma complementar, as diferentes perspectivas pelas quais as entidades do mundo real podem ser modeladas nos bancos de dados são apontadas por Ceri e Widom (1993) como ligadas a uma origem metodológica dessa diversidade semântica.

Já, sob um ponto de vista mais prático, Bright, Hurson e Pakzad (1994) afirmam que heterogeneidade semântica está ligada ao fato de que vários elementos de informação possuem diferentes nomes e estruturas nos diversos bancos de dados. Essa afirmação põe em evidência uma distinção que é característica presente em vários trabalhos. Neles há uma separação entre os problemas de nomenclatura e os ligados a estrutura com que os dados são armazenados internamente nos bancos de dados. Esta última categoria de problemas é, normalmente, classificada como heterogeneidade sintática e não semântica. No marcante artigo de Hull (1997), ele define que a heterogeneidade semântica é "o fato de dados duplicados através de múltiplos bancos de dados serem representados diferentemente no que há de subjacente aos esquemas de banco de dados". Destaca, também, que isso ocorre em um contexto onde diferentes conceitualizações relacionadas às informações a serem armazenadas são utilizadas para representá-las, além dos diferentes esquemas de banco de dados.

De forma mais detalhada, Hakimpour e Timpf (2001) afirmam que *semântica* está relacionado ao significado dos dados, colocando em contraste com a *sintaxe* adotada para armazená-los, isto é, a estrutura dos esquemas de bancos de dados. Ressalta que tal como há divergência entre as comunidades científicas sobre terminologia e conceitos usados para descrição dos assuntos estudados, também há diferentes formas de se interpretar os dados coletados e armazenados. Segundo eles, são essas diferentes interpretações dos dados que dão origem à heterogeneidade semântica, sendo que a dependência dessa significação implícita dos dados constitui a principal causa dessa diversidade observada nas bases.

Nessa linha, Mitra e Wiederhold (2002) notam que mesmo usando o mesmo formato de dados, fontes de informações podem utilizar estrutura e semântica que divergem. Afirmam, também, que "tal heterogeneidade é resultado da natureza autônoma das ontologias e do fato de que as fontes de informação são construídas por diferentes pessoas com diferentes objetivos em mente". Já Hellweg et al. (2011) observam que, no início de nossa década, "os usuários de serviços de informação estão encarando fontes de documentos altamente descentralizadas e heterogêneas, com diferentes análises de conteúdo". Segundo eles, uma diversidade semântica ocorre quando "recursos usando diferentes sistemas para

descrição de conteúdo são buscados usando um único sistema de consultas". Não deixam de enfatizar que a heterogeneidade semântica é muito mais complicada de solucionar do que uma heterogeneidade de tecnologias empregadas para prover esses serviços. Outra boa contribuição é feita por Horsburgh et al. (2014) ao enxergarem que a diversidade semântica dificulta a descoberta, integração e síntese dos dados e que isso se deve, muitas vezes, ao comportamento dos pesquisadores que usam diferentes termos para descreverem conceitos e à sua discordância a respeito deles. O que torna mais grave a situação é que eles raramente compartilharam os dados observados com as devidas e suficientes anotações e metadados. As devidas anotações fariam com que fosse possível para outro pesquisador interpretá-lo sem ambiguidades. Um interessante resumo do significado desse termo é trazido por Singh (2013), ao dizer que a heterogeneidade semântica ocorre "devido à diferença no conteúdo e seus significados subjacentes, isto é, ou o mesmo dado tem diferentes significados ou diferentes dados tem o mesmo significado".

2.2.2 Tipos e classificação

Além desses aspectos essenciais destacados, a literatura procura desenvolver algumas classificações dos tipos de heterogeneidade semântica. Isso é feito para melhor compreensão do tema e para construção de soluções para partes específicas do problema. Os autores Ventrone e Heiler (1991) fizeram um detalhado trabalho ao identificar sete situações de "evolução de domínio" no qual há, segundo eles, o surgimento de tipos de heterogeneidade semântica correspondentes à essas mudanças. Os tipos de "evolução de domínio" e os tipos de heterogeneidade semântica são plenamente análogos, pois o que aqueles descrevem sob uma perspectiva temporal também ocorre, segundo os mesmos autores, em um ambiente de combinação de múltiplos bancos de dados. Isso nos aponta que as categorias de diversidade semântica encontradas estão ligadas às situações estáticas e dinâmicas, estruturais e temporais, à própria natureza da heterogeneidade semântica e não à apenas um de seus aspectos. A classificação desenvolvida por eles é apresentada de forma sumarizada na Tabela 2.

No campo da comunidade de Geoinformática, outra abordagem é realizada por Worboys e Deen (1991) ao agruparem em duas grandes categorias os problemas de diversidade semântica. Eles afirmam que em um sistema distribuído espacialmente pode haver dois tipos de heterogeneidade semântica. A primeira, denominada de heterogeneidade semântica genérica, ocorre quando os nós que compõe um sistema estão usando diferentes modelos de informação espacial. A segunda, chamada de heterogeneidade semântica contextual, está ligada aos problemas resultantes das diferenças de modelagens nos bancos de dados geográficos no processo de integração das bases.

Já Ceri e Widom (1993) apontam quatro formas em que a heterogeneidade semântica pode se apresentar. A primeira ocorre nos conflitos de nomenclatura, isto é, quando são

Tabela 2: Tabela dos tipos de heterogeneidades semânticas.

Tipo de Heterogeneidade Semântica	Descrição
Instâncias heterogênicas	Diferentes ocorrências do mesmo valor podem possuir significados diferentes
Cardinalidade	Relacionamentos entre entidades podem mudar de acordo com o tempo ou domínio
Granularidade	Valores podem representar diferentes granularidades de um mesmo tipo de informação
Codificação	Valores podem ter sido armazenados de forma codificada, de modo que um mesmo código pode indicar vários significados
Tempo e unidades	Valores podem ter sido representados em diferentes instantes ou com diferentes unidades
Identificadores	Uma informação pode ser indexada por identificadores diversos
Reutilização de campos	Um campo que originalmente possui um significado é usado para outro propósito

Fonte: Segundo Ventrone e Heiler (1991).

armazenados diversos nomes que se referem ao um mesmo conceito; a segunda no que eles chamam de conflitos de domínios, que sucede quando diferentes bancos de dados se valem de valores distintos para indicar um mesmo conceito; a terceira forma de se apresentar são os conflitos de metadados, que aparece quando um mesmo conceito é representado em níveis distintos nos bancos de dados, por exemplo, quando se representa um conceito com uma instância em uma base e em outra se representa como um esquema; a quarta e última forma são nos conflitos estruturais, que acontecem quando os mesmos conceitos são representados usando formas de organização dos dados diferentes, isso em termos de modelagem conceitual, lógica e física. Hull (1997) comenta que, sob a perspectiva lógica ou semântica, a heterogeneidade estava presente na representação dos dados, que poderia se apresentar como divergência nos modelos de dados, esquemas e tipos de dados.

Mais uma outra classificação dos tipos de heterogeneidade é feita por Hakimpour e Geppert (2001), eles a separam em dois grupos. O primeiro tipo é chamado de heterogeneidade dos dados e consiste nas diferenças entre "definições locais, tais como tipos de atributos, formatos ou precisão"; e o segundo, é propriamente a heterogeneidade semântica, que eles definem como se referindo "as diferenças ou similaridades entre os significados dos dados locais". Mitra e Wiederhold (2002) realizam uma abordagem muito semelhante. É interessante ressaltar que essas classificações serviram de base para algumas etapas no trabalho realizado no Estudo de Caso proposto.

2.2.3 Onde ocorre

A heterogeneidade semântica é um fenômeno que ocorre em diversos ambientes. Para melhor ilustrar esse acontecimento, no contexto dos vários segmentos e ciências que se utilizam da Ciência da Computação e suas tecnologias, segue algumas situações que receberam atenção da literatura.

Abstraindo as características específicas de cada segmento é possível, em um primeiro momento, agrupar os problemas relacionados à heterogeneidade semântica em duas situações chave. A primeira é a ligada ao desenvolvimento temporal do significado dos dados em determinado domínio da ciência, que pode fazer com que os dados armazenados em um único banco de dados, por exemplo, não tenham o mesmo significado que tinham quando foram adicionados à base. Ventrone e Heiler (1991) põe em relevo que essa evolução de domínio "pode criar problemas de heterogeneidade semântica dentro de um [único] banco de dados semelhantes àqueles encontrados em sistemas de múltiplos bancos de dados" e que por isso requerem soluções semelhantes.

A segunda situação, mais fácil de notar, é a relacionada à distribuição espacial e de domínio dos dados armazenados. Foi devido à explosão do número de bancos de dados nos anos 80 que se tornou imperativo o desenvolvimento de soluções para problemas que surgiam no processo de interoperabilidade dos dados armazenados. No contexto de armazenamento de dados, a heterogeneidade semântica aparece com mais força nesse cenário. Bright, Hurson e Pakzad (1994) descreve essa situação da seguinte forma:

> (...) diferentes sistemas gerenciadores de bancos de dados (SGBDs), que geralmente são incompatíveis uns com os outros, foram desenvolvidos para ir de encontro com as variadas necessidades desses ambientes independentes. Contudo, no mundo conectado de hoje, fontes de dados autônomas e separadas, "ilhas de informação" [Andrew 1987], não são mais capazes de ir de encontro às necessidades cada vez mais sofisticadas dos usuários. Informações afins importantes para uma aplicação global ou requisição podem existir em múltiplas e incompatíveis bases locais. (BRIGHT; HURSON; PAKZAD, 1994)

Kashyap e Sheth (1998) notaram que "a escala do problema mudou de poucas bases de dados para milhões de recursos de informações". Já Colomb (1997) destaca que a dificuldade em desenvolver sistemas que facilitem a interoperabilidade entre recursos de informação se dá porque a demanda por isso surge justamente da presença de heterogeneidade semântica entre os esquemas e ontologias utilizados pelos diferentes serviços. Bergamaschi, Castano e Vincini (1999) complementam afirmando que a heterogeneidade semântica ocorre em larga escala envolvendo, entre outras coisas, os aspectos geográficos, organizacionais e funcionais relacionados ao uso da informação.

Voltando-se para os segmentos específicos, vários são os contextos onde se nota a presença da heterogeneidade semântica. Entre eles, o trabalho cooperativo, a fabricação

baseada em computadores e o processamento de dados. A heterogeneidade semântica também é comum em aplicações de domínio, tais como os sistemas de informação de escritório, os sistemas de fabricação integrada por computador (como design auxiliado por computadores), na computação pessoal, na computação empresarial e financeira, em bases de informação de pesquisas científicas, além de nos inúmeros serviços presentes na Web. Essa lista de ambientes nos dá uma visão do quão abrangente é a área de atuação desse problema, e de quantas são as pessoas interessadas em resolvê-lo (HAMMER; MCLEOD, 1993; FANG; HAMMER; MCLEOD, 1994; SHVAIKO et al., 2005).

Uma ênfase especial pode ser dada aqui à presença da heterogeneidade semântica no contexto da comunidade científica, tema desta dissertação. No início dos anos 90 já se tinha, claramente, que a abundância de dados coletados, associada às novas capacidades de armazenamento e processamento, geraria um movimento de especialização ainda mais acentuado em relação aos recortes de trabalho realizados por pesquisadores. Já se notava que isso exigiria uma conduta ainda mais voltada para o compartilhamento eficiente de informações e conhecimentos adquiridos ao longo das pesquisas científicas. É nesse sentido que Hammer e McLeod (1993) advertem que:

> (...) a qualidade e progresso dos esforços científicos dependem da habilidade dos pesquisadores de eficientemente armazenar grandes quantidades de dados heterogêneos e, ainda mais importante, compartilhar e trocar esse conhecimento com seus/suas colegas. (HAMMER; MCLEOD, 1993).

Na senda dessas atividades, vários pesquisadores se dedicaram em seus campos específicos por apresentar contribuições e propostas de soluções para os problemas que enfrentavam relacionados à heterogeneidade semântica. É possível citar os esforços na comunidade de Geoinformática, como o de Bishr (1998). Ainda na Geoinformática Hakimpour e Timpf (2001) apontam para a necessidade de Sistemas de Informações Geográficas (SIGs) interoperáveis. Eles dizem isso devido ao crescente aumento de dados geográficos, além de destacarem que esses sistemas não devem apenas lidar com a dificuldade de "compartilhar e integrar dados entre sistemas com diferentes modelos e estrutura de dados, ela tem também que lidar com a heterogeneidade semântica". Dizem ainda que isso se tornou mais importante, pois o foco da modelagem de dados espaciais está voltado para os diferentes modelos de dados presentes em pesquisas e no mercado. Na mesma área, Singh (2013) recorda que os sistemas de dados espaciais gerenciam grande quantidade de dados de fontes e formatos heterogêneos.

Outra área de grande importância tem sido a Web Semântica. Gracia e Mena (2012) explicam que ela se apresenta como uma gigante fonte de informações, de forma distribuída e heterogênea, além de aumentar o seu tamanho muito rapidamente devido ao grande número de usuários que fazem uso dela adicionando novas informações.

Na comunidade de hidrologia, Horsburgh et al. (2014) afirmam que heterogeneidade semântica continua a ser uma das dificuldades na descoberta e integração de dados. Eles também destacam que "a habilidade de descobrir e integrar dados de múltiplas fontes, projetos, e esforços de pesquisa é crítico para que os cientistas continuem a investigar processos hidrológicos complexos ao expandir escalas espaciais e temporais".

2.2.4 Soluções

Na busca de soluções para os diversos problemas relacionados à heterogeneidade semântica surgiram várias abordagens, dentre as quais algumas merecem ser destacadas, ao menos em linhas gerais. É possível observar certa evolução histórica no desenvolvimento e adoção das técnicas, como segue. No início dos anos 90, Ventrone e Heiler (1991) argumenta que a maneira mais efetiva de lidar com os problemas de heterogeneidade semântica é transformá-los quando possível em problemas sintáticos, "fazendo a semântica explícita nos dados e na aplicação". Dessa forma poderiam ser mais facilmente tratados através da manipulação da sintaxe desses elementos. Em outra perspectiva Ceri e Widom (1993) chamam a atenção para algumas funcionalidades presentes em alguns ambientes que lidam com múltiplos bancos de dados. Destacam que as regras de produção e pilhas de persistência poderiam ser usadas para gerenciar e manter a consistência semântica entre vários bancos de dados.

Ainda no âmbito dos banco de dados Hammer e McLeod (1993) propõem um *framework* e uma arquitetura que promovem uma abordagem que se adapta a uma federação de banco de dados interoperáveis, autônomos e heterogêneos. Usando para isso um modelo de dados comum e funcionalidades que identificam relacionamentos entre objetos de informação. Subindo um pouco nível de abstração, Bright, Hurson e Pakzad (1994) desenvolveram o *Summary Schemas Model* (SSM), Modelo de Sumário de Esquemas em tradução livre, que consiste em uma descrição concisa e mais abstrata do conteúdo semântico de um grupo de esquemas. O SSM "usa relacionamentos linguísticos específicos entre os termos do esquema para construir uma estrutura de dados global e hierárquica que descreve as informações disponíveis em todos os bancos de dados locais de forma cada vez mais abstrata". Com outra contribuição Fang, Hammer e McLeod (1994) fazem menção ao projeto da *University of Southern California* (USC) chamado *Remote Exchange* que procurava inventar, e implementar experimentalmente, técnicas e mecanismos para um compartilhamento controlado de informações.

Explorando um pouco mais as soluções da heterogeneidade semântica no contexto dos bancos de dados, Kashyap e Sheth (1998) lista três grandes abordagens para lidar com múltiplos bancos de dados. São elas: a federação fortemente acoplada, federação fracamente acoplada e gerenciamento de dados interdependentes. Com relação à federação fortemente acoplada, Colomb (1997) analisa a heterogeneidade semântica nesses sistemas

com objetivo de ver a viabilidade de acrescentar a interoperabilidade no nível semântico a esse tipo de federação. Ele chega à conclusão de que ela provavelmente estará presente se os elementos participantes da federação não trocarem mensagens entre si, ou todos não fornecerem uma descrição padronização para um ambiente comum. Sob uma ótica de mais alto nível, Aslan e McLeod (1999) utilizaram um modelo canônico de dados detalhado para lidar com o problema. Eles também recordam que uma das abordagens possíveis para mitigar o problema da heterogeneidade semântica consiste em ontologias ou dicionários semânticos em que

> (...) os participantes concordam em relação a um conjunto de conceitos do mundo real e relacionamentos entre esses conceitos. Cada banco de dados participante é responsável por expressar a porção compartilhável de seu esquema conceitual nos termos do vocabulário comum. O compartilhamento e troca de informações ocorre pela análise dos conceitos vigentes indicados pelos elementos dos bancos de dados individuais, pela investigação dos relacionamentos entre conceitos, e pela derivação dos significados de conceitos desconhecidos quando necessário. (ASLAN; MCLEOD, 1999)

O trabalho de Bergamaschi, Castano e Vincini (1999) já utiliza tesauros com o intuito de compartilhar uma ontologia das fontes de informações.

Afastando-se um pouco das soluções voltadas somente para o nível de banco de dados e adotando técnicas mais generalistas, Hakimpour e Geppert (2001) apresentam uma abordagem que verifica similaridades em ontologias formais. Essa verificação é realizada em um determinado domínio e usa essas detalhes semânticos para resolver o problema da heterogeneidade em um esquema global. De forma complementar, Hakimpour e Timpf (2001) mostram que ontologias podem ajudar as aplicações a deixarem de serem tão dependentes desses significados implícitos dos dados, que eles chamam de "conhecimento de fundo" da comunidade em questão. Afirmam, também, que um dos objetivos de tornar esse conhecimento explícito através de ontologias é descobrir as "discrepâncias nas extensões dos conceitos". Shvaiko et al. (2005) já adicionam suporte à semântica a um sistema de *Match*. Fazem isso para que ao usar a infraestrutura de Redes de Inferências fosse demonstrado que era possível explorar o conteúdo semântico dos dados nos resultados. Dentro da comunidade hidrologia, Horsburgh et al. (2014) compartilha um vocabulário e ferramentas de *software* para incentivar e demonstrar a importância de se compartilhar a semântica dos dados a fim de "alcançar um consenso a respeito da linguagem de compartilhamento que pode ser usada para descrever dados de observação hidrológica". Nesse mesmo caminho, Singh (2013) utiliza ontologias em um *framework* que abrange a modelagem e as consultas em sistemas de dados espaciais.

Os conteúdos semânticos dos registros ficaram durante muito tempo implícitos dentro das comunidades de uso daquelas informações. Os anos 90 foram marcados por diversas propostas que buscavam retirar os significados dos dados desses segmentos de

pessoas e torná-los públicos ao registrá-los sob diversas técnicas. Nos anos 2000, a adição de metadados sobre os registros e o uso de ontologias para organização global desses conceitos foram consolidadas como as abordagens mais comuns para lidar com o problema de heterogeneidade semântica.

2.3 Considerações sobre o Capítulo

Como se pode ver, a literatura nos mostra o interessante desafio que a Ciência moderna encontra ao lançar mão de um novo paradigma metodológico. À modelagem matemática se uniram os modelos computacionais, e com eles todas as técnicas e infraestruturas necessárias para se dar suporte a eles, em termos computacionais e de interação entre os pesquisadores. Como destaca Candela et al. (2015), os dados obtidos pelas mais diversas fontes possuem uma estrutura interna e de armazenamento heterogênea. Essa situação leva a necessidade de uma atribuição semântica as diferenças encontradas. Isso deve ocorrer para permitir o recorte e demais pré-processamentos necessários para que esses dados possam ser adequadamente comparados e utilizados. Naturalmente, essas atividades consomem tempo e recursos dos diversos projetos de pesquisa. As *e-Infrastructures* se mostram como uma solução tecnológica que permite tanto o suporte ao armazenamento desses dados como a sua reutilização. Além de fornecerem ferramentas utilizadas em pesquisas, como modelos e *software*. Entre elas, foram citadas as consideradas de maior pertinência para esse trabalho. Destacou-se a iniciativa do *D4Science*, e um projeto que a desenvolve e utiliza, o *EUBrazilOpenBio*, cujo foco é a biodiversidade em plantas. A título de referência, existe ainda o projeto *iMarine*, que utiliza o *D4Science*, porém, com foco em espécies marinhas.

Nesse capítulo se buscou fazer um apanhado sumário dos conceitos fundamentais ligados ao aparecimento da Computação Científica como novo paradigma de metodologia científica, bem como as consequências provenientes de sua utilização. Por fim, se mostrou algumas soluções que lidam nesse ambiente de Computação Científica com os problemas inerentes a heterogeneidade de dados e a interoperabilidade de ferramentas e informações entre esse tipo de comunidades científicas.

***

3 Materiais e Métodos

Este trabalho está inserido dentro de um contexto de produção científica realizada por algumas linhas e grupos de pesquisa que atuam no Programa de Pós-Graduação de Física Ambiental e no Instituto de Computação. Esse contexto é análogo ao de tantos outros locais envolvidos com ciências que lidam com dados ambientais. Como já mencionado o intuito mais abrangente deste trabalho ao abordar o problema da diversidade semântica é superá-lo ou mitigá-lo a fim de que os dados ambientais possam ser devidamente utilizados para os fins acadêmicos. Além de ajudar na integração e cooperação entre as diversas partes envolvidas nesses ambientes e projetos.

Este capítulo busca trazer de forma sumarizada os materiais (seção 3.1) e metodologias (seção 3.2) adotadas para alcançar os objetivos deste trabalho. Nas seções abaixo são elencados os principais componentes da solução proposta nesta dissertação para mitigar o problema da heterogeneidade semântica dos dados ambientais de um contexto igual ou semelhante ao PGFA.

3.1 Materiais

Os materiais escolhidos para utilização neste trabalho foram selecionados com base em: (a) sua adequação aos objetivos do trabalho e (b) sua maior facilidade de manutenção. Para atender ao segundo critério foram adotados preferencialmente *software* de licença aberta. Isso proporciona estar inserido dentro de uma grande comunidade desenvolvedores que trabalham para correção, adequação e atualização constantes das ferramentas. Além de fóruns e ambientes de discussão, que são de grande ajuda para dar suporte à configuração e customização das ferramentas.

3.1.1 gCube

É possível descrever sucintamente o que levou a adoção do *gCube* como material para este trabalho da seguinte forma. Como já foi mencionado (subseção 2.1.2.3), o *gCube* é uma tecnologia que pode ser descrita sob três principais pontos de vista, como um provedor de uma infraestrutura de dados, um *framework* para dados e processamento ou uma aplicação científica. Uma descrição geral que sintetiza essas realidades é a que o entende como "um modo de se entregar aplicações científicas na nuvem." (GCUBE CONSORTIUM, 2015).Desse modo, entre vantagens encontradas, pode-se ressaltar que ele permite a integração, o compartilhamento e a cooperação no armazenamento e uso dos dados e metadados. Permite o mesmo também em relação às ferramentas e aos

algoritmos de processamento aplicados sobre esses conjuntos de dados. Isso é feito por meio da implementação de conceitos de organização de infraestruturas virtuais amplamente explorados, como VO, VRE e Objetos de Pesquisa. Ressalta-se também que o *gCube* foi escolhido por ser um *framework* sob uma licença aberta que provê infraestruturas virtuais capazes de disponibilizar uma grande gama de ferramentas que permitem oferecer como serviços recursos computacionais para o uso de grupos de pesquisa. Atentou-se para o suporte que ele dá ao uso de metadados e a validação por meios desses metadados, que atuam para diminuição do problema de heterogeneidade semântica, fazendo com os dados se tornem semanticamente viáveis para outros grupos. Outro fator de escolha é a estabilidade do projeto, visto que possui grandes parceiros que lhe dão maior confiabilidade. Em especial, é mantido em colaboração com órgãos da União Europeia.

3.1.2 D4Science

Para adotar o *gCube* como ferramenta foi escolhida uma implantação concreta em larga escala, que já está em uso por diversas iniciativas. Trata-se da infraestrutura virtual *D4Science*, que é mantida por uma organização de mesmo nome. Essa organização oferece uma HDI, como destacado anteriormente (subseção 2.1.2.4), para diversos colaboradores do âmbito científico e com o uso da tecnologia *gCube*.

O *D4Science* foi escolhido ao invés do uso direto da tecnologia *gCube* devido à alguns fatores. Essas características podem ser observadas na parte inferior Tabela 1. O *D4Science* oferece alguns recursos a mais, desenvolvidos para o uso em colaboração com o *gCube*. Algo a se ressaltar são os serviços oferecidos pelo *D4Science*. Sem custos em relação a infraestrutura virtual é possível criar e/ou utilizar os diversos serviços que um Ambiente Virtual de Pesquisa (VRE) pode oferecer. Nesse ambiente é possível também ter acesso às diversas aplicações criadas por eles para pesquisa científica. Também é possível implantar algum serviço ou aplicação desenvolvida localmente. Como está exposto posteriormente (seção 4.1), essa possibilidade foi explorada. Algo que também está a disposição são os conjuntos de dados organizados pelos colaborados do grupo ou mesmo de outros VREs. Há também nesse cenário todas as ferramentas necessárias para a gestão e coordenação do acesso e papéis das pessoas do grupo (D4SCIENCE INFRASTRUCTURE, 2015c).

Pode-se destacar também que o *D4Science* oferece uma gama de diversos componentes de *software* e integra virtualmente mais de 50 provedores de dados, além de todas as funcionalidades e serviços disponíveis no *gCube*. Isso tudo dentro de um ambiente coerente fornecido pela infraestrutura virtual. É interessante destacar que a escolha dessa solução tecnológica para o uso do *framework gCube* se deu também pela reduzida alocação de recursos para seu uso; de fato, para realização de todas atividades deste trabalho não foi requisitada nenhuma taxa de serviço. Evitou-se também os infortúnios envolvendo a manutenção do *hardware*, da plataforma e demais *software* necessários para manter uma

infraestrutura própria. Uma contribuição direta para esse critério de escolha, foi o fato de terem sido empenhadas várias horas deste trabalho na tentativa de implementar uma infraestrutura virtual baseada em *gCube* em *hardware* local. Essa atividade se demonstrou demasiada onerosa e por isso foi descartada. Depois disso, esta pesquisa usufruiu de todas as facilidades que a abordagem em Nuvem do *D4Science* proporciona (D4SCIENCE INFRASTRUCTURE, 2015b).

3.2 Métodos

Na realização deste trabalho naturalmente algumas escolhas metodológicas foram feitas. Elas foram tomadas na busca de abordar adequadamente o assunto e alcançar os objetivos propostos. Entre elas é possível ressaltar as que seguem. Em relação ao tipo de dissertação realizada, foi escolhida a dissertação monográfica, que por sua natureza se ocupa de um assunto específico, e busca expor o tema de modo educativo e com a metodologia adequada. Pode-se dizer, também, que é uma dissertação expositiva, pois foi feita de modo a reunir nesse trabalho o material essencial a respeito do assunto pesquisado, utilizando-se de um gama diversa de fontes, e buscando expor de forma organiza e com exatidão as descrições dadas pelos autores a respeito do assunto, com seus problemas e das soluções que emergem desse meio (MARCONI; LAKATOS, 2003).

Com relação à abordagem se trata de uma pesquisa qualitativa, que não busca traduzir numericamente os enunciados para cumprir seus objetivos, utilizando-se amplamente da descrição como ferramenta para avaliar e validar a adequação da solução adotada ao objetivo do trabalho. Pela natureza do objetivo do trabalho que, como foi definido anteriormente, passa majoritariamente pela demonstração da mitigação da heterogeneidade semântica dos dados ambientais comumente utilizados no contexto do Programa de Pós-Graduação de Física Ambiental, julgou-se adequado a utilização desse método, pois contempla a exposição do Estudo de Caso que mostra a realização e cumprimento desse mesmo objetivo (LAKATOS; MARCONI, 1986).

Depois dessa descrição geral da metodologia adotada, e antes de detalhar os caminhos tomados em partes específicas do trabalho, é útil por em evidência a pergunta de pesquisa de norteou todo este trabalho. Durante ele, e principalmente ao final, buscou-se responder a seguinte questão: *uma infraestrutura virtual distribuída como o D4Science é capaz de fornecer os meios para mitigar ou superar os problemas de heterogeneidade semânticas que surgem no processo de produção científica em uma ciência que lida com dados ambientais?*

Em relação às partes mais específicas, pode-se começar destacando algumas escolhas feitas a respeito dos critérios da pesquisa bibliográfica. Foram escolhidos dois principais assuntos, a Heterogeneidade Semântica e a Computação Científica. Para o primeiro assunto,

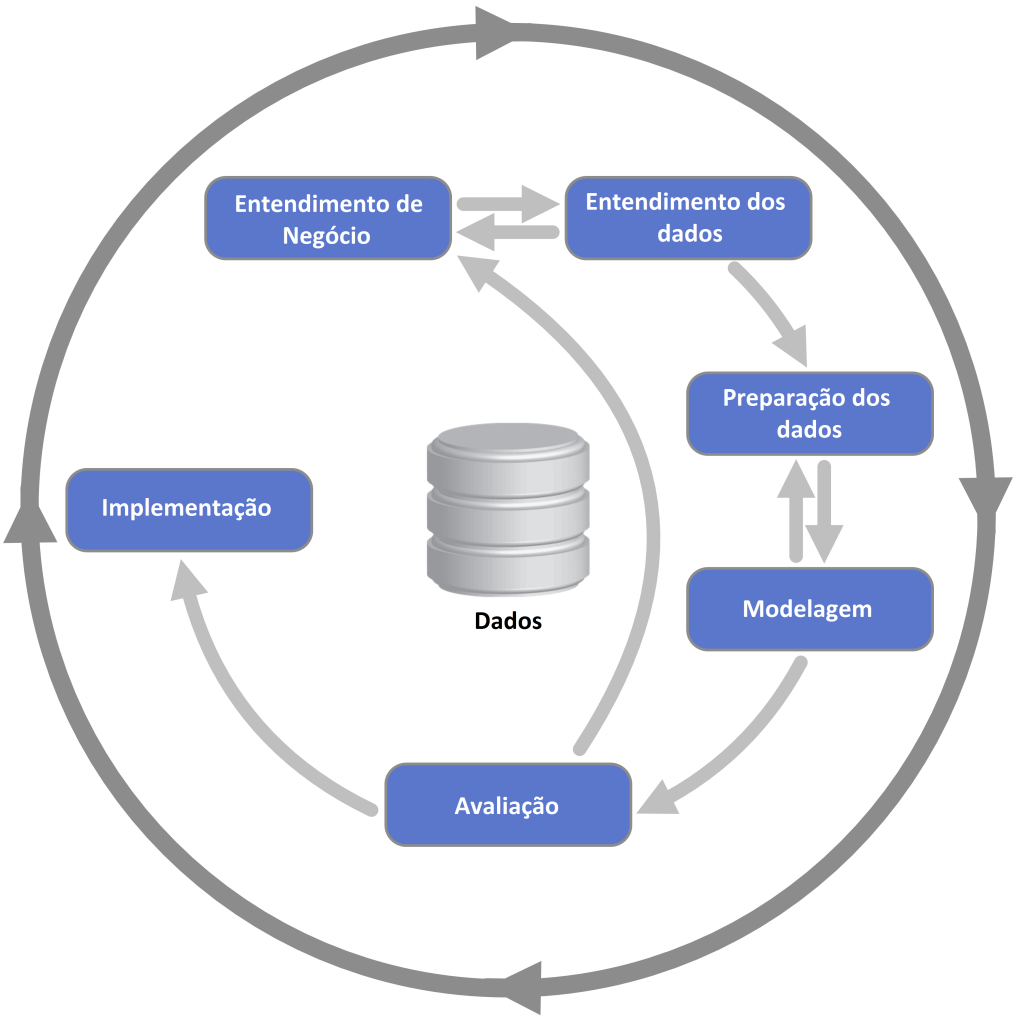
adotou-se uma palavra-chave *semantic heterogeneity*. Os artigos foram coletados em dois grupos, o primeiro grupo reuniu os artigos até 2010. O segundo grupo coligiu os artigos após 2010. Os dois grupos continham vários artigos, mas no final cada um permaneceu com cerca de 20 artigos coletados entre os de maior relevância segundo a indexação de motores de busca científicos como *Google Scholar* e o Portal de Periódicos CAPES/MEC.

Para o assunto da Computação Científica foram utilizadas quatro palavras-chave *Scientific Computing*, *Cloud Computing*, *e-Infrastructure* e *Hybrid Data Infrastruture*. Na coleta desses artigos foram escolhidos os trabalhos de mais relevância de acordo com os indexadores dos motores de busca *Google Scholar* e o Portal de Periódicos CAPES/MEC, sem nenhuma restrição de datas. Inicialmente foram escolhidos vários trabalhos, porém, ao final de uma seleção que levou em conta a afinidade dessas pesquisas com os objetivos deste trabalho, permaneceram por volta de 10 à 15 artigos relacionados a cada palavra-chave.

É interessante colocar em ênfase alguns pontos sobre o Estudo de Caso desenvolvido para este trabalho. Ele foi elaborado para se explorar a capacidade da infraestrutura virtual escolhida em dar resposta aos problemas de heterogeneidade semântica comumente encontrados no ambiente de pesquisa da Física Ambiental. Nesse sentido, ele permite encontrar os elementos para a resposta da pergunta de pesquisa exposta anteriormente. Para realização deste Estudo de Caso elaborou-se algumas tarefas. Elas podem ser vistas como um detalhamento de trabalhos que envolvem as fases de Entendimento dos Dados, Preparação dos Dados e Modelagem do *Cross Industry Standard Process for Data Mining* (CRISP-DM). Esse padrão é um modelo de processos usado largamente na comunidade mineração de dados. Na Figura 1 se encontra as principais fases desse modelo para melhor compreensão da natureza transversal das tarefas realizadas no Estudo de Caso (IBM CORPORATION, 2011).

Essas tarefas foram agrupadas em um roteiro usado para a mitigação da heterogeneidade semântica em dados científicos. O seu conteúdo será descrito no Capítulo 4 por meio de diagramas de *Business Process Management* (BPM) e explicações a respeito de sua aplicação.

Figura 1: Figura que mostra as fases do Modelo de Processos CRISP-DM



Fonte: Baseado em Wikimedia Commons (2015) e IBM Corporation (2011)

3.3 Considerações sobre o Capítulo

Neste capítulo foi exposto de forma sumarizada os materiais e as escolhas metodológicas adotadas para este trabalho. Primeiramente, ressaltou-se o contexto em que a pesquisa foi realizada, dentro do Programa de Pós-Graduação em Física Ambiental e com o apoio do Instituto de Computação na UFMT. Em seguida, voltou-se atenção para os critérios de escolha dos materiais utilizados, que priorizaram o *software* livre e aberto, devido a sua maior facilidade de manutenção e replicação do ambiente criado nesse trabalho para pesquisadores futuramente interessados.

Os principais materiais foram elencados, posteriormente, junto de suas características mais relevantes e com destaques às funcionalidades adotadas para realização deste trabalho. O *gCube* foi apresentado como uma ferramenta aberta que permite a implementação de infraestruturas virtuais. Em seguida, foram feitas considerações a respeito das

razões que levaram a escolha dessa tecnologia. Em um segundo momento, deu-se destaque ao *D4Science*, uma organização que disponibiliza uma infraestrutura virtual baseada na tecnologia *gCube*. Ela dispõe de uma infraestrutura para dados híbridos, que se mostra adequada aos propósitos deste trabalho.

Finalizando as discriminações necessárias, foram elencadas as escolhas metodológicas adotadas. Fez-se notar que este trabalho é de natureza monográfica e expositiva, e que utiliza uma abordagem qualitativa para cumprir seus objetivos. Explicou-se as práticas adotadas para seleção do material da revisão bibliográfica. Por fim, foram apresentadas as motivações para o Estudo de Caso e a existência de um roteiro de atividades seguidas dentro dele.

O próximo capítulo apresenta a exposição das condições de realização do Estudo de Caso, as funcionalidades adotadas nas ferramentas escolhidas, a descrição dos resultados obtidos após a validação do uso dos metadados propostos pela infraestrutura virtual e a discussão sobre esse uso como meio de mitigação da diversidade semântica dos dados.

***

4 Resultados e Discussão

Este trabalho tem por finalidade expôr como a adoção de infraestruturas virtuais e seus serviços podem contribuir para mitigação da heterogeneidade semântica presente em dados ambientais. Nesse sentido, este capítulo se preocupou em adotar um itinerário que vai desde a identificação da diversidade semântica em dados de bases ambientais até a constatação de que a adoção de algumas técnicas específicas desses ambientes virtuais foram eficazes ao oferecer soluções para os tipos diversidade semântica presentes neste Estudo de Caso.

Para melhor compreensão, ele foi divido segundo a seguinte estrutura. Primeiramente se elaborou um estudo de caso (seção 4.1), que contempla algumas fases necessárias para a realização do objetivo deste trabalho. Depois, em um segundo momento (seção 4.2), é feita a discussão à respeito de como se deu a mitigação da heterogeneidade semântica, junto de considerações sobre os benefícios trazidos pelo uso de infraestruturas virtuais, tanto na mitigação do problema tratado como em outras circunstâncias comuns na Computação Científica.

4.1 Estudo de Caso

A realização deste Estudo de Caso foi levado a termo para a validação da assertiva de que o uso de infraestruturas virtuais pode constituir uma ferramenta eficaz para tratar o problema de heterogeneidade semântica presente nos dados científicos e, no contexto da Física Ambiental, especificamente nos dados ambientais. Vale notar que a adoção de técnicas para contornar a diversidade semântica dos dados é uma área de intensa pesquisa por parte da comunidade que investiga a evolução da Computação Científica.

Para evidenciar que determina técnica foi bem sucedida ao tratar determinados conjuntos de dados com algum tipo de heterogeneidade semântica, é preciso que, ao final dos procedimentos, os conjuntos de dados que inicialmente apresentavam incompatibilidades de natureza semântica possam ser utilizados como uma fonte unificada de dados para processamentos e/ou outras inferências.

No intuito de realizar precisamente essa validação conceitual foi escolhido um itinerário, que compõe os tópicos desta seção. Nesse sentido, para deixar mais claro os passos adotados durante a execução deste Estudo de Caso, um roteiro do tratamento dos dados é apresentado a seguir. É importante notar que as etapas finais dele servem como uma avaliação se a solução implementada foi capaz de mitigar os problemas de heterogeneidade semântica encontrados nas situações propostas.

O roteiro do tratamento da diversidade semântica nos dados é o seguinte:

a) definição e configuração de um VRE que ofereça serviços para mitigar a inconsistência semântica dos dados;

b) para cada situação proposta dentro do Estudo de Caso as seguintes etapas são desenvolvidas:
   - descrição específica da situação proposta, sua natureza e finalidade;
   - detalhamento dos conjuntos de dados escolhidos para aquela situação, sua origem, natureza e abrangência;
   - a realização de uma classificação dos tipos de heterogeneidade semântica encontrados nos conjuntos de dados da situação proposta;
   - exposição específica da adoção de um esquema de metadados para validar a integração semântica dos conjuntos de dados que possuem heterogeneidade semântica;
   - descrição das transformações feitas nos conjuntos de dados pelos serviços do VRE para alcançar a homogeneidade semântica definida no esquema de metadados;
   - validação da homogeneidade semântica dos conjuntos de dados por meio da aplicação do modelo de metadados especificado no esquema;
   - mescla de forma indistinta dos dados dos conjuntos de dados que inicialmente eram heterogêneos e que agora obedecem o mesmo esquema de metadados;
   - validação da homogeneidade dos dados por meio do processamento indistinto deles por um algoritmo incorporado ao VRE.

Essas etapas são sintetizadas graficamente por meio do Diagrama de Atividades ilustrado na Figura 2.

A tarefa de *criar um modelo de metadados* presente na Figura 2 é de especial importância para este trabalho. Por isso, descrevê-la como um processo que agrupa outras subtarefas pode dar uma melhor visão, com maior detalhamento, de como ela foi realizada na proposta deste Estudo de Caso. Ela pode ser esquematizada como mostrado na Figura 3.

Em um primeiro passo é pedido a criação de um novo modelo de dados. Essa tarefa no *D4Science* pode ser realizada por meio da funcionalidade de criação de *template* presente no Serviço *Tabular Data Manager* (TabMan). A próxima tarefa é a definição dentro desse modelo da descrição da dimensão temporal da série temporal. Isso pode ser feito na mesma funcionalidade do TabMan, basta que se defina uma variável como sendo do tipo *TIMEDIMENSION*, se escolha a resolução temporal desejada e o formato que deve ser utilizado. Em seguida, deve-se definir variáveis que serão atributos da série. Para isso, é necessário definir uma ou mais variáveis com o tipo *ATTRIBUTE*, os tipos de dados dessas variáveis e o formato como elas serão representadas. Depois disso se pode definir

Figura 2: Figura que mostra o Roteiro para mitigar a heterogeneidade semântica

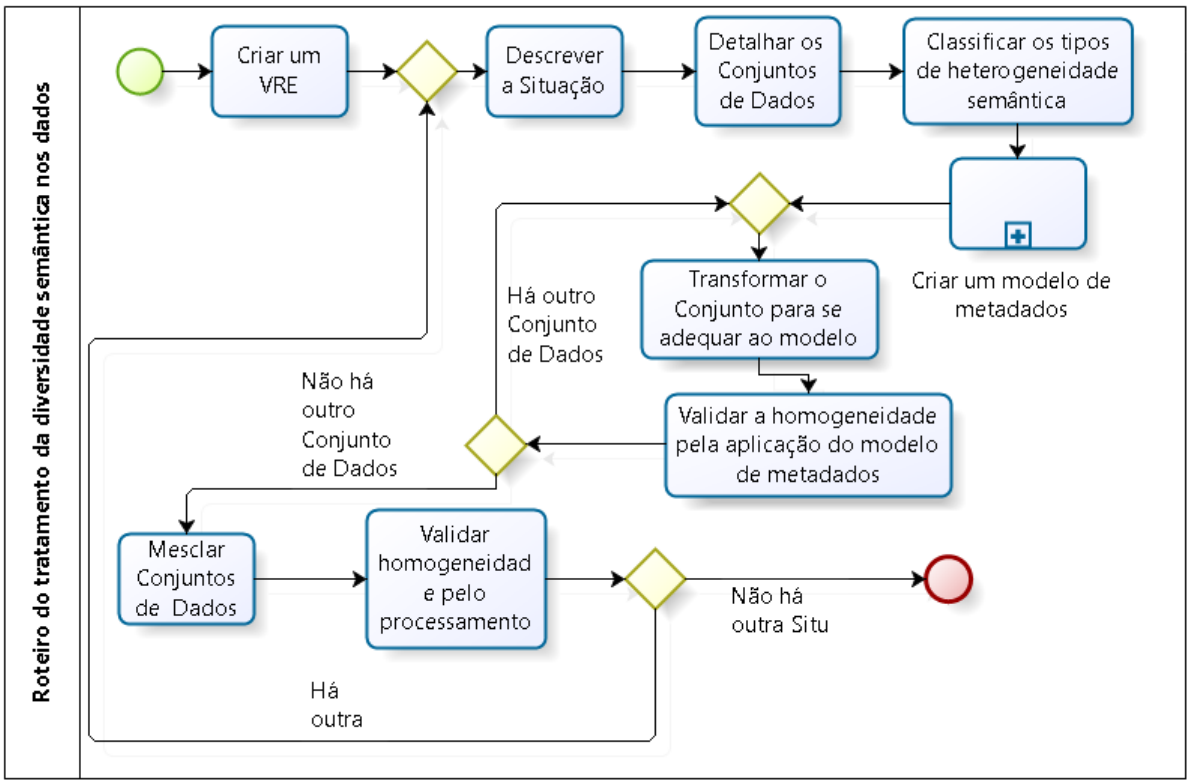
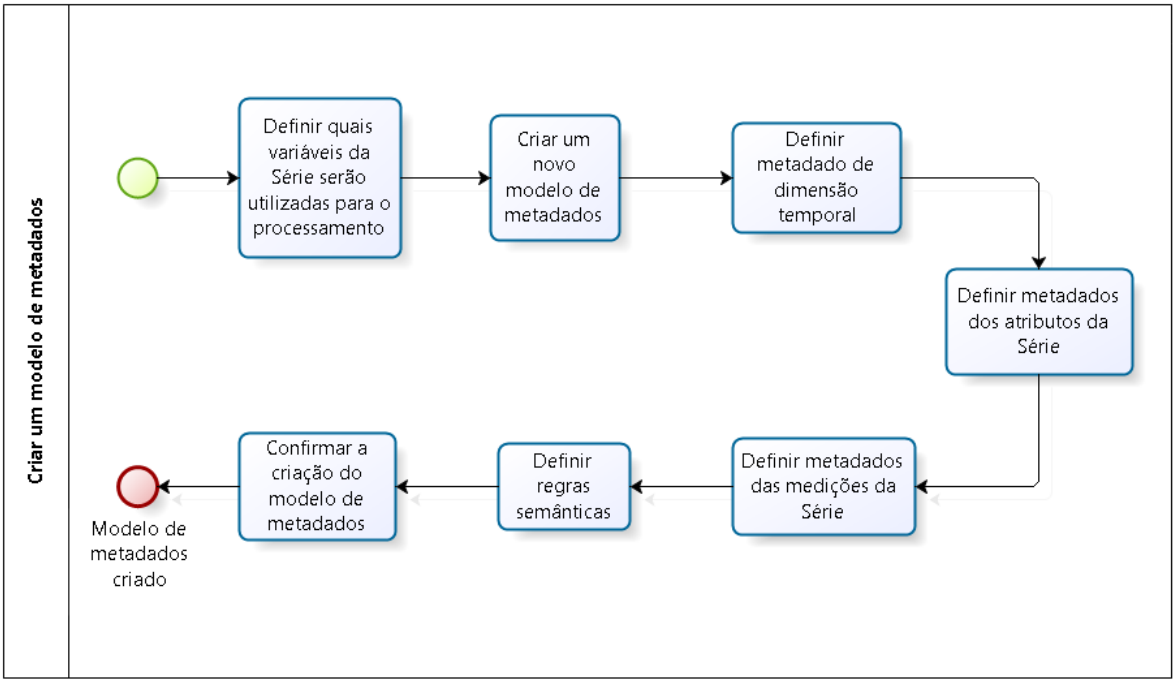


Figura 3: Figura que mostra o processo de criação de um modelo de metadados



as variáveis que propriamente são as medições do conjunto de dados. A fim de realizar essa tarefa se deve definir uma ou mais variáveis com o tipo *MEASURE*, os tipos de dados dessas variáveis e o formato adotado para representá-las. Por fim, se pode adicionar alguma restrição semântica por meio do criador de *templates* do TabMan. As restrições se baseiam na definição de condições para que os dados sejam admitidos em uma determinada dimensão temporal, atributo ou medição. Após essas subtarefas basta confirmar a criação do modelo de metadados para finalizar o processo.

4.1.0.1 D4Science e a Criação de um VRE

A primeira atividade que se deve fazer a fim de se utilizar a infraestrutura virtual *D4Science* é criar um conta para acessar seus serviços. Depois disso, para implementação deste Estudo de Caso, se optou pela criação de um VRE próprio dedicado a esse domínio de atividades ligadas a pesquisa com dados ambientais. O nome escolhido para esse Ambiente Virtual de Pesquisa foi PGFA-UFMT, como pode ser visto na tela inicial apresentada na Figura 4. Como já foi enfatizado pela literatura (cf. subseção 2.1.2) um VRE é um ambiente de pesquisa completo voltado para o armazenamento, gerenciamento, processamento e compartilhamento de dados e ferramentas de *software* ligadas a pesquisa científica. Foi dentro desse ambiente que foram realizadas as atividades relativas a Infraestrutura *D4Science*.

Entre os serviços do VRE existe um que dá suporte ao gerenciamento de recursos tabulares, ele é o já mencionado *Tabular Data Manager* (TabMan). Primeiramente se procedeu com a importação dos dados a partir do arquivos fornecidos pelas respectivas instituições. O processo de importação dos arquivos *.csv* ocorreu normalmente mesmo com séries temporais significativas. Uma vez que os dados foram importados para dentro do VRE o *TabMan* pode atuar sobre os dados por meios de diversas ferramentas que ele oferece, como será visto nas próximas seções.

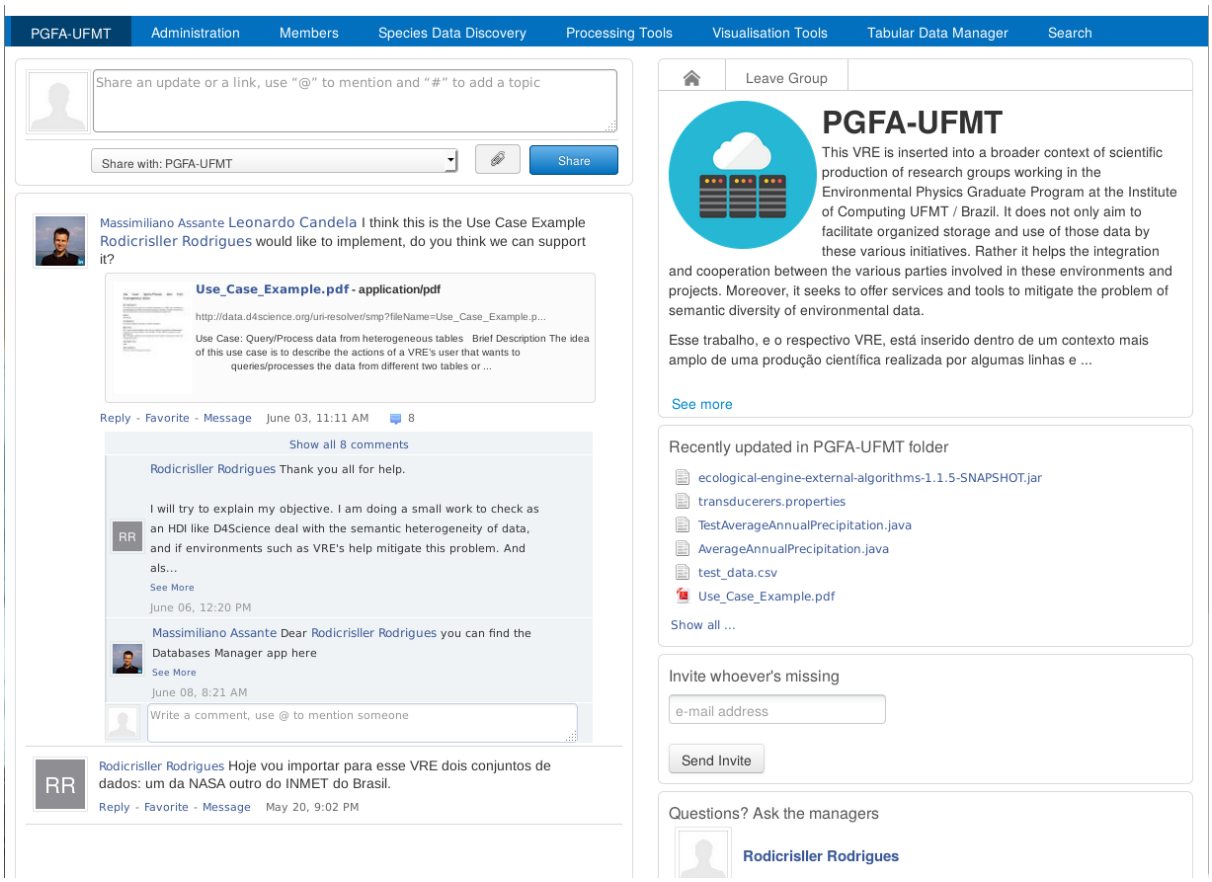
4.1.1 Situação I

Na realização deste Estudo de Caso foram propostas duas situações, que visam ilustrar atividades habituais dos envolvidos com o uso de dados ambientais. Esse é o caso dos estudantes e professores do PGFA e do IC. Na *Situação I* foram escolhidas duas fontes de dados que possuem, dentre outras variáveis, dados de *precipitação* referentes a região de Cuiabá/MT. A finalidade dessa escolha é, uma vez superada as dificuldades provenientes da diversidade semântica, processar os dados de modo a fornecer a *precipitação anual*.

4.1.1.1 Dados Ambientais

Com o objetivo de exemplificar um caso comum de obtenção de dados provenientes de fontes diversas, foram selecionados quatro conjuntos de dados dispostos dentro de duas

Figura 4: Figura com tela inicial do VRE PGFA-UFMT



Fonte: *D4Science Infrasctructure*

situações. Embora, nas duas situações, esses conjuntos de dados tratem do mesmo domínio de informação eles não apresentam uma plena coerência semântica.

É comum no campo das ciências ambientais que, embora tratem do mesmo assunto, e muitas vezes das mesmas variáveis micrometeorológicas, esses conjuntos de dados tragam algum tipo de heterogeneidade semântica. Isso ocorre devido a diversidade de técnicas e dispositivos utilizados para o sensoreamento de determina região. Em algumas situações as diferenças nas formas de registro das informações não aparecem apenas pela diferença nos gêneros de sensores, mas também por se tratarem de fabricantes distintos que não buscam padronizar a saída de dados.

Essa situação acrescenta certo trabalho ao especialista e, em alguns casos, essa tarefa de harmonização dos dados não encontra uma maneira adequada de ser padronizada. E mesmo nas vezes que se consegue realizá-la, geralmente, não é feita por meio de uma ferramenta que possa ser compartilhada e reutilizada de forma adequada pelos pesquisadores posteriormente.

Os dois primeiros conjuntos de dados que se referem à *Situação I*, por fins práticos

são denominados *Conjunto A* e *Conjunto B*. Os dados do *Conjunto A* são provenientes de sensoriamento remoto, e foram obtidos por meio da *National Aeronautics and Space Administration* (NASA), Administração Nacional da Aeronáutica e do Espaço em tradução livre [1]. Foram selecionados os dados de *precipitação* da fonte com a descrição de *Daily TRMM 3B42 V7 Rainfall (3B42 V7)*. Para o presente trabalho os dados foram recortados, como será descrito posteriormente, para área que compreende a região de Cuiabá/MT. Além da *precipitação* os registros trazem consigo as informações de: data, latitude, longitude e o índice do registro. O Conjunto A possui dados de 1999 à 2009.

Outros dados de *precipitação* foram obtidos para o *Conjunto B* dos registros do Instituto Nacional de Meteorologia (INMET), de uma estação automática de Cuiabá/MT, de mnemônico Cuiabá-A901, com código na Organização Mundial de Meteorologia (OMM) nº 86705, latitude de -15.559295º, longitude de -56.062951º e altitude de 242 metros. Os registros selecionados foram feitos de 2010 à 2013. Os dados desse conjunto dizem respeito à: código da estação, data, hora, temperatura do ar (instantânea, máxima e mínima), umidade do ar (instantânea, máxima e mínima), ponto de orvalho (instantâneo, máximo e mínimo), pressão atmosférica (instantânea, máxima e mínima), vento (velocidade, direção, rajada), radiação e precipitação.

4.1.1.2 Classificação dos Tipos de Heterogeneidade Semântica

Os dados obtidos, como ocorre frequentemente, apesar de serem da mesma cidade e domínio apresentam diversidade semântica em sua composição. Para ilustrar quais são os tipos de heterogeneidade semântica presentes nos dados segue uma exposição comparativa dos dois conjuntos de dados selecionados.

Na realização dessa análise e classificação se levou em conta os critérios apontados pela literatura (cf. subseção 2.2.2) como os mais relevantes. Principalmente as classificações propostas por Ventrone e Heiler (1991) e Ceri e Widom (1993) em seus trabalhos.

Segundo os diversos autores adotados, em um primeiro momento deve-se diferenciar a heterogeneidade dita *sintática* da propriamente *semântica*. Nesse sentido, não serão feitas considerações a respeito da disposição dos dados dentro das estruturas adotadas para armazená-los. Abaixo são apresentados na Figura 5 e na Figura 6 os cabeçalhos dos conjuntos de dados *A* e *B*.

[1] <http://disc.sci.gsfc.nasa.gov/services/grads-gds>

Figura 5: Figura com o cabeçalho do *Conjunto A* de dados

datatrmm	lat	lon	precipitac	id
1999-01-01	-22.125000000...	-61.875000000...	0.000000000000...	942712
1999-01-01	-22.125000000...	-61.625000000...	0.000000000000...	942713
1999-01-01	-22.125000000...	-61.375000000...	0.000000000000...	942714
1999-01-01	-22.125000000...	-61.125000000...	0.000000000000...	942715

Fonte: *NASA*

Figura 6: Figura com o cabeçalho do *Conjunto B* de dados

codigo_estacao	data	hora	temp_inst	temp_max	temp_min	umid_inst	umid_max	umid_min	pto_orvalho_inst	pto_orvalho_max	pto_orvalho_min	pressao
A901	01/01/2010	00	25.4	25.5	25.1	77	79	77	21.1	21.4	21.1	990.9
A901	01/01/2010	01	25.2	25.5	24.8	76	79	76	20.8	21.4	20.8	991.8
A901	01/01/2010	02	25.3	25.4	24.4	75	87	75	20.5	22.1	20.5	992.5
A901	01/01/2010	03	24.7	25.7	24.7	78	78	72	20.6	20.7	20.1	991.9

Fonte: *INMET*

Uma vez feita essa ressalva e desconsiderando esse tipo de heterogeneidade nos dados se pode seguir rumo a identificação dos tipos de inconsistências semânticas para cada situação. Uma análise qualitativa buscando aplicar aos conjuntos de dados os conceitos de heterogeneidade semântica propostos pela literatura pode ser descrita como segue.

Levando em consideração os dados dos *Conjuntos A e B*, que estão presentes na *Situação I*, destacam-se as seguintes categorias de diversidade semântica:

a) o que se nota já em um primeiro momento é uma heterogeneidade semântica de ***tempo***, uma vez que a resolução temporal dos dois conjuntos de dados não estão com a mesma periodicidade. Os dados do *Conjunto A* estão com uma periodicidade diária, enquanto os registros do *Conjunto B* foram armazenados a cada hora;

b) um segundo tipo de diversidade semântica encontrada é a de ***identificadores***. O *Conjunto A* possui um campo que indexa as informações por uma numeração que não carrega nenhum sentido dentro do recorte de dados efetuado, menos ainda em relação ao outro conjunto de dados. Já o *Conjunto B* possui um identificador mnemônico para a estação de origem dos dados, informação essa que servia para indexar esses dados em uma base de maior abrangência, mas que não guarda significado para o contexto atual dos dois conjuntos de dados;

c) o que se destaca também é o ***conflito de nomenclatura***. Isso ocorre frequentemente em conjuntos de dados de origens distintas. Nesse caso, essa circunstância foi observada em todos os campos dos conjuntos de dados.

4.1.1.3 Definição de Template com Metadados

Efetuado o levantamento dos tipos de heterogeneidade semântica presentes nos conjuntos de dados se pode dar seguimento às atividades que buscam efetivamente eliminar, senão ao menos mitigar, a presença dessas inconsistências de ordem semântica em vista de um processamento unificado dos dados.

Uma das melhores soluções apresentadas pelos pesquisadores dessa área é fazer com que o conhecimento a respeito dos dados – isto é, seus metadados – estejam explícitos, de modo que possam ser compartilhados e utilizados (cf. subseção 2.2.4). Nesse sentido, o *TabMan* oferece uma ferramenta capaz de criar, gerenciar, compartilhar e aplicar *templates* em conjuntos de dados. Esses *templates* conseguem expressar certo conteúdo semântico dos dados, o que faz com que os dados que passam por sua validação possuam uma consistência sintática e semântica entre si.

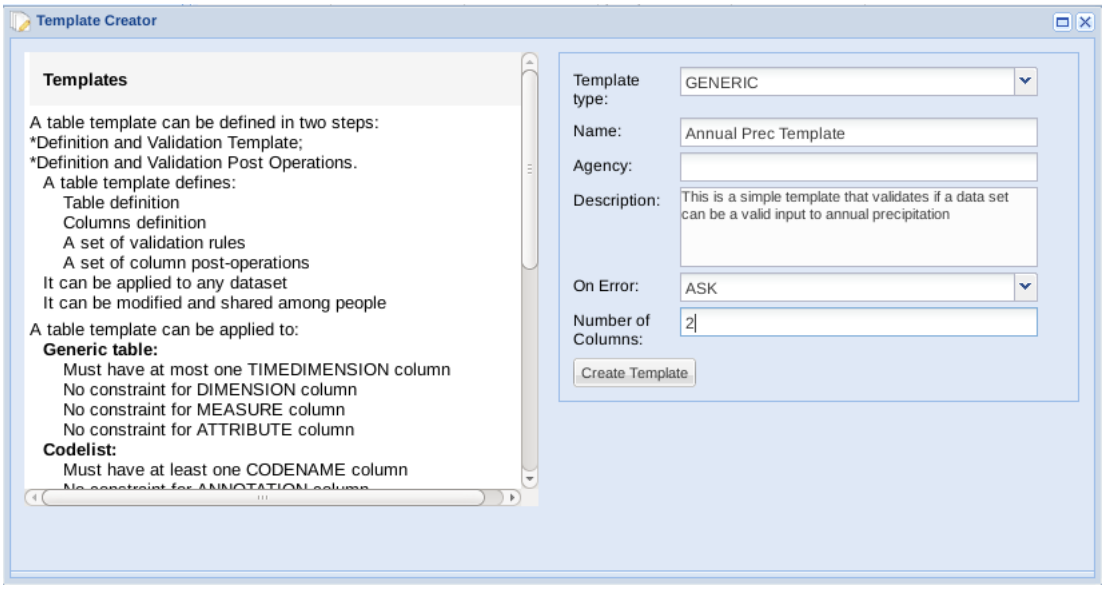
Buscando alcançar uma coesão semântica entre os conjuntos de dados da *Sitação I*, foi elaborado um *template* de metadados. Ele reúne em si as informações mínimas para

que esses conjuntos de dados possam ser utilizados como entradas válidas para o algoritmo de validação desenvolvido para cumprir a etapa final deste Estudo de Caso.

Tendo em consideração a *Situação I* foi criado um *template* que agrupe os dados de *data* e *precipitação*. Essas foram consideras as informações mínimas para o processamento que obterá a precipitação anual da série temporal oriunda da mescla dos dados do *Conjunto A* e do *Conjunto B*.

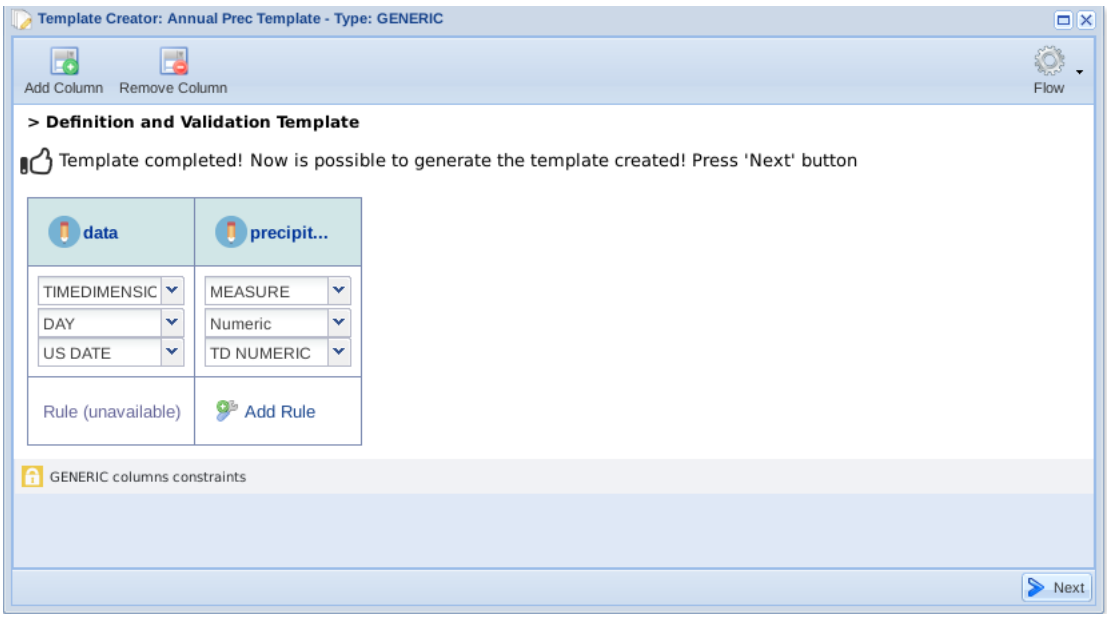
Uma descrição da elaboração do *template* da *Situação I* pode ser feita como segue. A Figura 7 mostra a etapa onde se define as informações básicas a respeito do *template*. Já na Figura 8 é mostrado o ambiente onde se definem os metadados do conjunto de dados. Como se pode notar, existe a opção de criação de campos com informações adicionais a respeito da natureza dos dados que eles irão armazenar. No caso, para definir o primeiro campo, que armazenará a *data* do conjunto de dados, foi escolhido o tipo de dados *TIMEDIMENSION* com periodicidade diária (*DAY*) e com formato americano de datas (*US DATE)*. Isso faz com que esse campo exija que o conjunto de dados inteiro possua uma resolução temporal diária. O segundo campo foi definido como sendo do tipo medida (*MEASURE*), com tipo de dados numérico (NUMERIC) e com formato de número decimal (*TD NUMERIC*). Essas escolhas definem que essa informação é a medida de uma variável, e não mais um atributo da tabela.

Figura 7: Figura com o criador de Template do TabMan



Fonte: *D4Science Infrasctructure*

Figura 8: Figura com o criador de Template do TabMan, na tela de configuração das variáveis



Fonte: *D4Science Infrasctructure*

4.1.1.4 Transformações de Dados

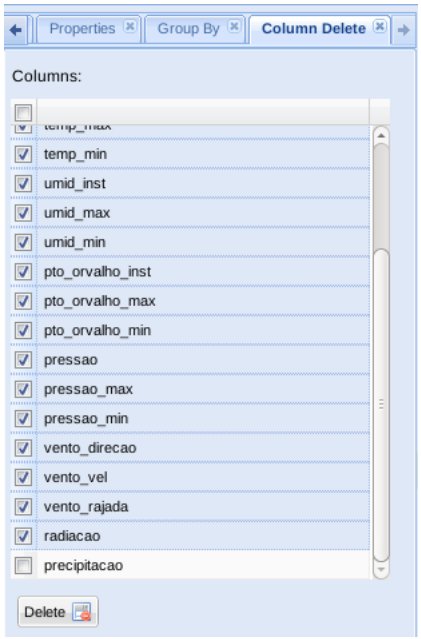
Definida uma descrição semanticamente homogênea por meio do *template*, o destino a se chegar a partir dos dados brutos fica claro. Para se conformar ao *template* os conjuntos de dados precisam passar por uma série de transformações para superarem cada um dos casos de heterogeneidade semântica identificados. Durante esse processo a capacidade da infraestrutura virtual *D4Science* de prover os meios necessários para essas transformações será o alvo de atenção. A seguir, para a *Situação I*, está descrita a sequência de procedimentos que foram necessários para solucionar os tipos de diversidade semântica presentes em cada conjunto de dados.

Na *Situação I* existem os dados do *Conjunto A* e do *Conjunto B* que precisam corresponder ao *template* gerado para esse caso. Isso deve acontecer a fim de não terem mais distinções semântica que os impeçam de serem processados pelo algoritmo que fornecerá a *precipitação anual* para Cuiabá/MT.

O primeiro procedimento que foi feito nessa situação consistiu na adequação dos conjuntos de dados a uma mesma semântica em relação ao ***tempo***. No *Conjunto A* não foi necessária nenhuma adequação, pois possuía a mesma resolução temporal que o *template* adotou. Já o *Conjunto B* sofreu a alteração de sua resolução temporal que se encontrava em termos de horas e passou a ser diária. Para preparar essa transformação foram eliminadas antes as informações que não serão utilizadas para os propósitos do processamento futuro. Essa eliminação foi realizada pela deleção de colunas do recurso tabular, como é exibido

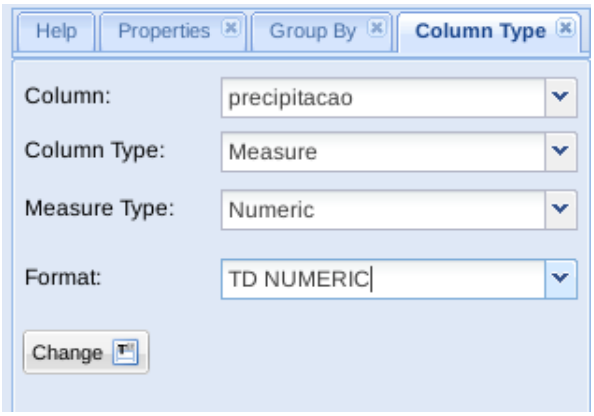
na Figura 9. Outra preparação foi a definição do campo *precipitacao* como um tipo medida numérica decimal, ela é mostrada na Figura 10. Depois dessa preparação, a etapa de adequação da resolução temporal foi feita graças a uma ferramenta de agregação de variáveis em função de outras variáveis, como é mostrada na Figura 11. Dessa forma o conjunto de dados passou a estar com uma periodicidade diária, como *template*.

Figura 9: Figura mostra a deleção de variáveis pelo TabMan



Fonte: *D4Science Infrasctructure*

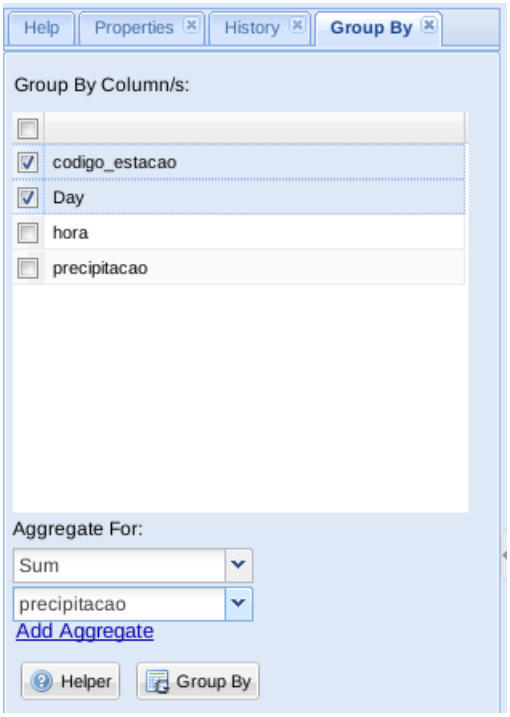
Figura 10: Figura mostra a definição de tipos de variáveis pelo TabMan



Fonte: *D4Science Infrasctructure*

Após esse procedimento o campo que representa a data do *Conjunto B* foi adequado ao formato americano de datas. Isso se fez necessário, pois essa formatação será cobrada

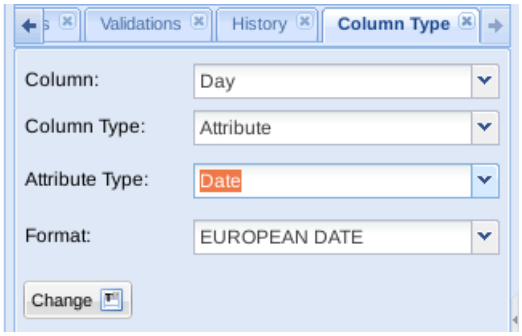
Figura 11: Figura mostra a mudança de resolução temporal pelo TabMan



Fonte: *D4Science Infrasctructure*

pelo *template* feito para a *Situação I*. É de fazer nota que essa operação se mostrou bem menos prática que as outras dentro do VRE. Foi necessário definir a data com o formato europeu no qual ela estava, como mostra a Figura 12. E depois disso voltar a defini-la como texto, pois o *template* trabalha com os dados brutos, como é o caso de um texto, para converter uma variável em *TIMEDIMESION*.

Figura 12: Figura mostra a mudança do tipo de dados da data do Conjunto B pelo TabMan



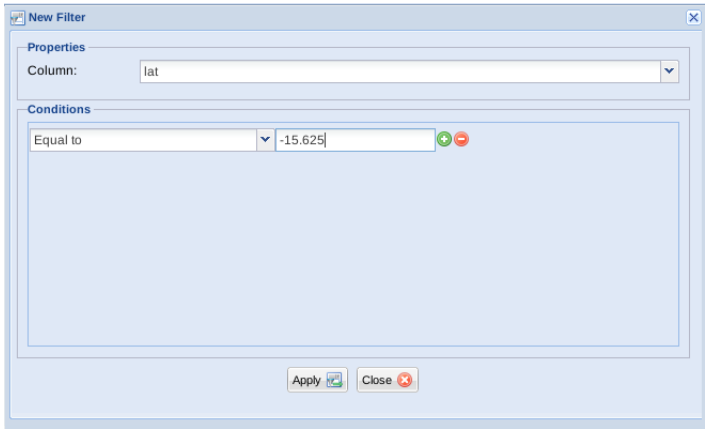
Fonte: *D4Science Infrasctructure*

Em seguida foi realizado o procedimento de adequação semântica em relação aos ***identificadores***. Para isso foi necessário a remoção dos campos *id* e *codigo_estacao* dos

Conjuntos A e B, respectivamente. Isso foi feito pela mesma ferramenta de deleção já citada. Outras providências não precisaram ser tomadas, pois o *TabMan* automaticamente gera novos índices unificados para os recursos tabulares quando eles são unificados, como acontecerá.

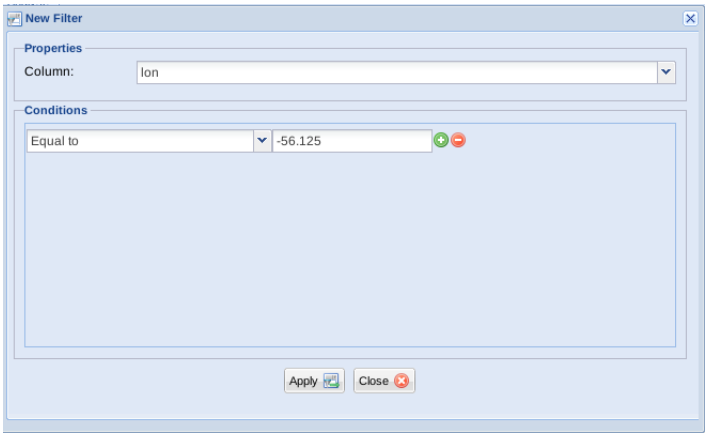
Embora não seja para resolver um tipo heterogeneidade semântica, um novo procedimento foi necessário devido a diferença de escopo espacial dos dados. Como foi dito o problema reside no *Conjunto A*, que possui uma abrangência espacial de uma granularidade maior que o do *Conjunto B*. Para corrigir essa inconsistência do *Conjunto A* em relação ao *Conjunto B* e ao *template* adotado, foi feito um recorte nos dados. O recorte foi feito de modo a obter os dados referentes a região mais próxima de a posição do *Conjunto B* (latitude de -15.559295º e longitude de -56.062951º), como mostram a Figura 13 e a Figura 14.

Figura 13: Figura o recorte da latitude do Conjunto A pelo TabMan



Fonte: *D4Science Infrasctructure*

Figura 14: Figura mostra o recorte da longitude do Conjunto A pelo TabMan

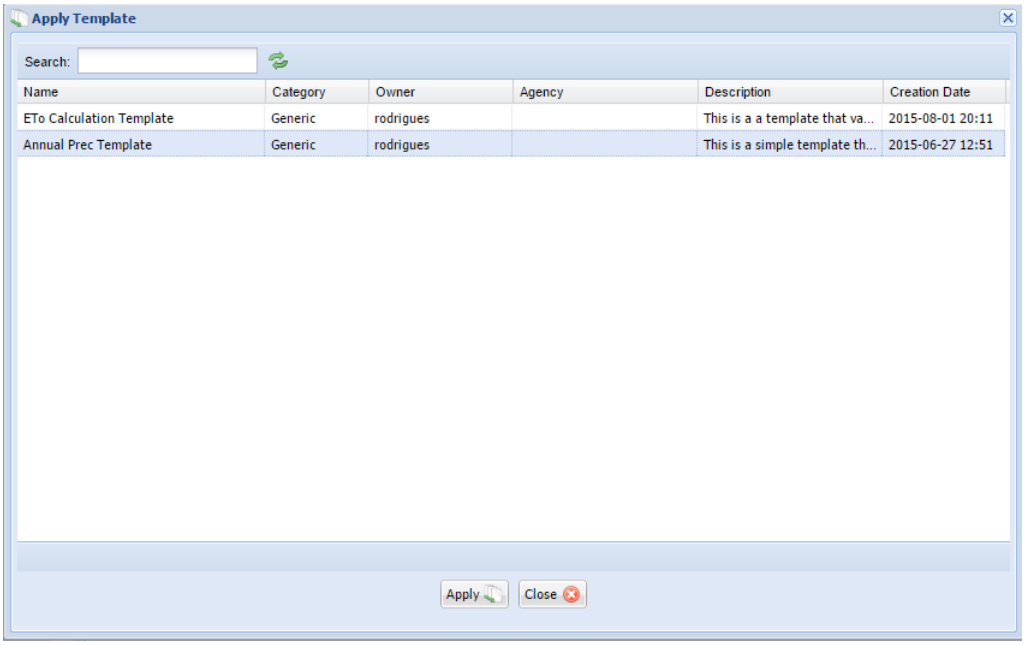


Fonte: *D4Science Infrasctructure*

Com relação a heterogeneidade semântica classificada pela literatura como ***conflito de nomenclatura***, também foi necessária uma ação prévia a aplicação do *template*. Isso ocorreu usando a ferramenta de edição de rótulos, que foi de fácil utilização.

Por fim, para realizar a completa conformação dos conjuntos de dados ao *template*, os dados de ambos os conjuntos foram submetidos a aplicação do *template* criado para a *Situação I*. Isso foi feito como ilustra a Figura 15. Durante essa aplicação o *TabMan* aplica a todos os campos a configuração e formato dos respectivos tipos de dados que foram definidos para cada um deles no *template*. A aplicação dessa técnica faz com que os dois conjuntos obedeçam a uma estrita definição sintática e semântica dos dados.

Figura 15: Figura mostra TabMan aplicando o template ao Conjunto A



Fonte: *D4Science Infrasctructure*

É de se ressaltar que existem outras formas de realizar transformações sofisticadas de dados. Por exemplo, uma delas é terceirizando-as para a *Statistical Manager* (StatMan), por meio da implementação de algum algorítimo específico que pode ser incorporado a infraestrutura posteriormente.

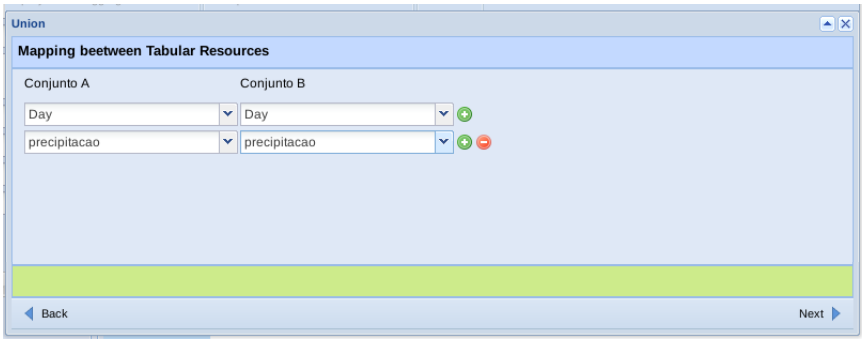
4.1.1.5 Mescla de Conjuntos de Dados

Os conjuntos de dados A e B que inicialmente apresentavam heterogeneidade semântica de diversos tipos foram submetidos a uma série de transformações. Esses procedimentos foram realizados por meio da infraestrutura virtual *D4Science* com o intuito de mostrar sua capacidade de superar esses vários pontos de diversidade semântica. Uma validação necessária para mostrar que os dados realmente não possuem mais nenhuma

das inconsistências semânticas que apresentavam é a plena compatibilidade na mesclagem desses dados e posterior processamento. A fim de que esses dados possam ser unificados, alguns procedimentos para a mesclagem das fontes de dados podem ser descritos como seguem.

Após a aplicação do *template* para *Situação I* os *Conjuntos A e B* foram submetidos ao processo de mesclagem de dados por meio da ferramenta *StaMan* do VRE PGFA-UFMT. Os campos foram mapeados e a operação gerou um *Conjunto A* com todas as informações dele e com as anteriormente pertenciam somente ao *Conjunto B*. Fazendo com que a série temporal passe ser contada de 1999 até 2013, totalizando 15 anos de dados coesos. Eles poderão ser processados como os pesquisadores desejarem, e com quem esse novo conjunto for compartilhado. O procedimento mencionado é ilustrado pela Figura 16.

Figura 16: Figura que mostra a mesclagem do Conjunto A e do Conjunto B pelo TabMan



Fonte: *D4Science Infrasctructure*

4.1.1.6 Validação por Processamento Incorporado

Foi vista como oportuna uma outra etapa para a validação dos dados. Eles se submeteram a um certo tratamento, também por meio do *D4Science*. Essa nova etapa consiste na execução de processamento sobre os dados que foram unificados pelo *TabMan* no VRE PGFA-UFMT.

O VRE possui um serviço que se chama StatMan, como já foi mencionado. Por meio desse serviço ele entrega a capacidade de pesquisadores realizarem processamentos diversos sobre um conjunto de dados. Há diversos algoritmos já presentes na biblioteca do serviço. Porém, se optou por se implementar algoritmos próprios. Essa escolha foi feita para se utilizar algoritmos que sejam de maior significância para o contexto da Física Ambiental. Como mencionado, para a primeira situação foi desenvolvido um algoritmo que fornece a precipitação anual a partir de um conjunto de informações com resolução temporal diária.

Com o objetivo de utilizar os algoritmos implementados pelos próprios pesquisadores, o *StatMan* oferece a possibilidade de implantá-los no ambiente do VRE. Isso foi realizado pela utilização de uma *Application Programming Interface* (API) em Java que o *gCube* fornece para essas situações. Para o desenvolvimento de um algoritmo há a necessidade de preencher os requisitos de uma interface que define os métodos mínimos de um algoritmo que será executado no *StatMan*. Essa interface é a *StandardLocalExternalAlgorithm*, que possuí um corpo mínimo de métodos como indica a Figura 17.

Figura 17: Figura que mostra a interface StandardLocalExternalAlgorithm

```
public class SimpleAlgorithm extends StandardLocalExternalAlgorithm{

	@Override
	public void init() throws Exception {
		// TODO Auto-generated method stub
	}
	@Override
	public String getDescription() {
		// TODO Auto-generated method stub
		return null;
	}
	@Override
	protected void process() throws Exception {
		// TODO Auto-generated method stub

	}
	@Override
	protected void setInputParameters() {
		// TODO Auto-generated method stub

	}
	@Override
	public void shutdown() {
		// TODO Auto-generated method stub
	}
	@Override
	public StatisticalType getOutput() {
		return null;
	}
}
```

Fonte: Segundo gCube Consortium (2015b)

É possível notar que a interface buscou reduzir ao máximo a necessidade de codificação extra por parte do especialista, permitindo que ele foque mais no algoritmo de seu processamento do que nos detalhes da API. O processo de implementação consiste basicamente em preencher as entradas pelo método *setInputParameters()*, as saídas pelo método *getOutput()* e o processamento propriamente dito pelo método *process()*.

Passando para os algoritmos desenvolvidos para a validação deste Estudo de Caso, o algoritmo de precipitação anual da *Situação I* pode ser descrito como segue. É importante mencionar que a implementação completa se encontra no Apêndice I. Foi escolhido como entrada um recurso tabular do próprio VRE. Essa entrada segue as especificações do *template* produzido para a primeira situação. Uma vez que um conjunto de dados passe pela validação do *template*, ele está apto a ser um entrada para esse algoritmo. A saída foi feita também por meio de uma recurso tabular no VRE, ela traz a precipitação total para cada ano que tenha algum valor nos dados de entrada. O processamento seguiu por meio da somatória das entradas levando em consideração o ano da coleta dos dados. Isso pode

ser observado pela algoritmo em linguagem natural exibido pela Figura 18.

Figura 18: Figura que mostra o algoritmo da Situação I em linguagem natural

```
leia listaDeDatas
leia listaDePrecipitacoes
real valorTotal
para i de 0 ate listaDeDatas.tamanho faca
  se i = 0 entao // e a primeira iteracao
    valorTotal <- listaDePrecipitacoes(i)
    se i = 1 entao // e a primeira e ultima iteracao
      escreva listaDeDatas(i), valorTotal
    fim se
  fim se
  se i > 0 entao
    data listaDeDatas(i)
    ultimaData listaDeDatas(i-1)
    se data.ano > ultimaData.ano entao
      escreva ultimaData.ano, valorTotal
      valorTotal <- listaDePrecipitacoes(i)
    senao
      valorTotal <- valorTotal + listaDePrecipitacoes(i)
    fim se
  se i <= 0 e i = listaDeDatas.tamanho // ultima interacao
    escreva listaDeDatas(i), valorTotal
  fim se
fim para
```

Após o processo de implementação, o código-fonte e o pacote de códigos compilados são submetidos aos administradores da infraestrutura que fazem a sua implantação. A interface gráfica para o algoritmo dentro do *StatMan* é gerada automaticamente por meio da API, que renderizada os componentes dentro do VRE conforme o tipo de entrada e saída escolhidas para o processamento. Isso é ilustrado pela Figura 19.

Depois da implantação do algoritmo, o Conjunto A de dados foi submetido ao processamento.

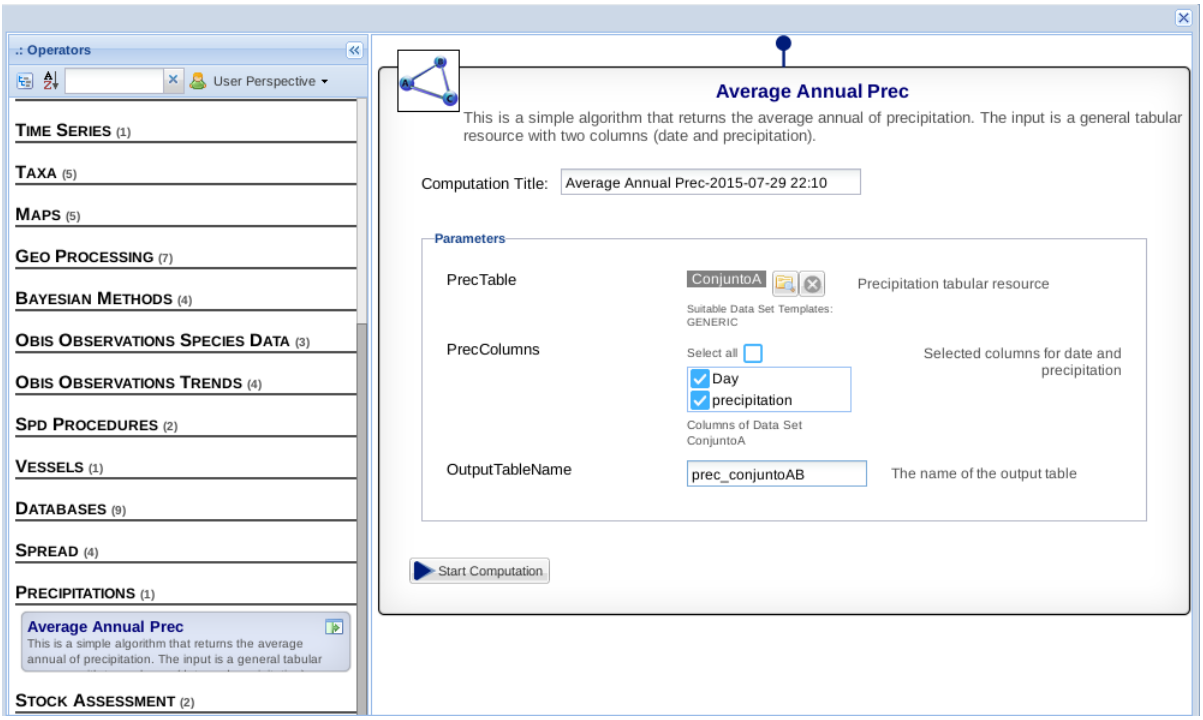
O resultado da *Situação I* pode ser observado na Tabela 3[2]. É possível notar em destaque a região do ano de 2010 até o ano de 2013, que correspondiam inicialmente ao Conjunto B, mas que cuja origem para o processamento final foi indiferente.

4.1.2 Situação II

A *Situação II* consiste na resolução dos problemas de heterogeneidade semântica de dois conjuntos de dados provenientes de duas estações meteorológicas, que estão situadas

[2] Os resultados dos anos de 2010 e 2011 apresentam uma diferença significativa em relação aos outros anos devido à uma falha na captação dos dados. Embora seja algo que deva ser tratado dentro da atividade de pré-processamento não apresenta relação com os problemas de heterogeneidade semântica. Isso faz com que não interfira nas conclusões dentro do recorte temático deste trabalho.

Figura 19: Figura que mostra a interface gráfica gerada pela API do gCube



Fonte: *D4Science Infrasctructure*

Tabela 3: Tabela do resultado do processamento da Situação I.

Ano	**Precipitação**
1999	1307.86
2000	1449.75
2001	1725.31
2002	1637.44
2003	1582.49
2004	1504.06
2005	1411.31
2006	1645.68
2007	1508.29
2008	1319.1
2009	1741.94
2010	**1056.4**
2011	**420.8**
2012	**1495.8**
2013	**1264.2**

no município de Cuiabá/MT. Isso será realizado a fim de fornecer informações coesas para o cálculo de evapotranspiração de referência (ETo) pelo método de Penman-Monteith-FAO. Para a chegar a esse fim e mitigar a diversidade semântica foram aplicadas as mesmas etapas da situação anterior, o que permite uma descrição mais sucinta nessa situação.

4.1.2.1 Dados Ambientais

Para a *Situação II* deste Estudo de Caso foram separados outros dois conjuntos de dados que serão denominados para fins práticos de *Conjunto C* e *Conjunto D*. Os dados do *conjunto C* foram obtidos também nos registros INMET por meio da estação meteorológica convencional de Cuiabá/MT, com código na OMM nº 83361, latitude de -15,619722º, longitude de -56,108333º e altitude de 145 metros. A abrangência temporal dos dados vai de 1998 até 2004. Esse conjunto de dados traz consigo as seguintes variáveis: estação, data, hora, precipitação, temperatura do ar (média, máxima e mínima), insolação, umidade relativa e velocidade do vento média.

O *Conjunto D* de dados foi retirado também dos registros do INMET, da mesma estação automática do *Conjunto B*, porém em períodos diferentes de tempo. Devido a isso os dados desse conjunto são estruturalmente os mesmos do *Conjunto B*, embora ainda preserve distinções semânticas em relação a esse. Os registros do Conjunto D vão de 2005 à 2010.

4.1.2.2 Classificação dos Tipos de Heterogeneidade Semântica

Primeiramente, para uma caracterização dos conjuntos de dados, são apresentados na Figura 20 e na Figura 21 os cabeçalhos dos conjuntos de dados *C* e *D*, respectivamente.

Figura 20: Figura com o cabeçalho do *Conjunto C* de dados

Data	Hora	Precipitacao	TempMaxima	TempMinima	Insolacao	Temp Comp Me...	Umidade Relati...	Velocidade do ...
01/01/1998	0000	0	37.9	0	11.4	29.92	64.25	1.466667
01/01/1998	1200	0.4	0	24.3		0	0	0
02/01/1998	0000	0	36.9	0	7.8	28.36	73.5	1.2
02/01/1998	1200	0	0	25.4		0	0	0

Fonte: *INMET*

Figura 21: Figura com o cabeçalho do *Conjunto D* de dados

codigo_estacao	data	hora	temp_inst	temp_max	temp_min	umid_inst	umid_max	umid_min	pto_orvalho_inst	pto_orvalho_max	pto_orvalho_min	pressao
A901	01/01/2010	00	25.4	25.5	25.1	77	79	77	21.1	21.4	21.1	990.9
A901	01/01/2010	01	25.2	25.5	24.8	76	79	76	20.8	21.4	20.8	991.8
A901	01/01/2010	02	25.3	25.4	24.4	75	87	75	20.5	22.1	20.5	992.5
A901	01/01/2010	03	24.7	25.7	24.7	78	78	72	20.6	20.7	20.1	991.9

Fonte: *INMET*

Para a realização da classificação das variedades de heterogeneidade semântica para a *Situação II* também foi elaborada uma análise qualitativa da questão. Esforçou-se por identificar nos conjuntos de dados os principais pontos sugeridos pela literatura.

Essa abordagem relação à *Situação II*, ao verificar os registros dos *Conjuntos C e D*, identificou os seguintes tipos de diversidade semântica que aparecem em relevo:

a) uma categoria de heterogeneidade semântica que se observa é a de ***identificadores***. Os *Conjuntos C e D* possuem campos que que indexam as informações pelo o código da estação de origem dos dados. O problema ocorre, pois o código presente no *Conjunto C* alude ao usado pela OMM, enquanto o que consta no *Conjunto D* reporta-se a uma enumeração interna do INMET;

b) percebe-se também uma diversidade semântica de ***tempo***, como notada na *Situação I*. Porém, aqui há um diferença de outra ordem, pois os dados das estações convencionais são coletados duas vezes ao dia, ao passo que nas estações automáticas os registros são realizados de hora em hora;

c) além de uma resolução temporal distinta, em relação aos dados de *umidade relativa* e *vento*, estão presentes nos dados da estação automática do *Conjunto D* os valores de mínima e máxima, no caso da umidade relativa, e direção e rajada, no caso do vento. Isso introduz entre os dois conjuntos de dados um heterogeneidade semântica que pode ser classificada como sendo de ***granularidade***. É possível dizer isso, pois a mesma variável está sendo vista por aspectos diferentes, uma vez que o acréscimo dessas características gera um maior detalhamento da variável;

d) outro tipo de diversidade semântica que surgiu foi em relação as ***unidades*** de medida. Isso ocorre no *Conjunto D* na variável de *radiação*, pois a medida instantânea da variável é feita em $\mathrm{W/m^2}$ e o *template* exige uma resolução diária dos dados, cuja unidade de medida para essa variável passa a ser em $\mathrm{MJ/m^2}$ dia;

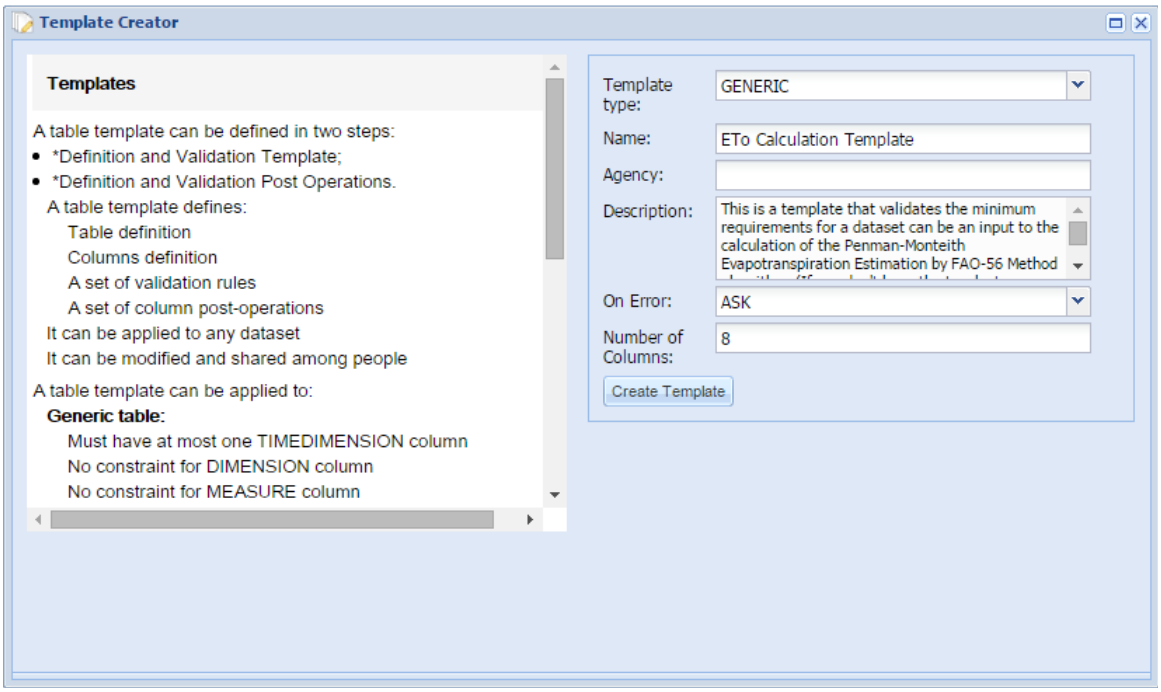
e) por fim deve fazer nota das comuns inconsistências entre rótulos de campos dos dois conjuntos de dados que gera uma diversidade semântica por ***conflito de nomenclatura***. É preciso mencionar que mesmo em dados provenientes da mesma instituição se pode encontrar esses casos de falta de padronização.

4.1.2.3 Definição de Template com Metadados

Depois de executado o levantamento dos tipos de heterogeneidade semântica presentes na *Situação II*, foi produzido um *template* que busca reunir as informações de entrada para o algoritmo de cálculo de evapotranspiração de referência. Esse *template* fará com que os dados que inicialmente possuem as heterogeneidades semântica apontadas

anteriormente passem a ter coesão semântica. O que permitirá que sejam utilizados indistintamente para o processamento pretendido. As variáveis de entrada que foram julgadas necessárias seguiram os apontamento de Conceição (2006), que descreve os procedimentos do Método de Penman-Monteith-FAO. A Figura 22 exibe a definição das informações de controle e identificação do *template*, de uso do *TabMan*. Já a Figura 23 ilustra a definição dos campos do *template*, que são associados a alguns metadados. Em ordem, foi definido primeiramente uma variável temporal (*TIMEDIMENSION*), com resolução diária (*DAY*) e em formato americano (*US DATE*). Em seguida, é mostrado os atributos (*ATTRIBUTE*) espaciais dos dados: a *latitude* e a *altitude*. Ambos do tipo numérico (*NUMERIC*), sob o formato de número decimal (*TD NUMERIC*). Após isso, configurou-se metadados para as variáveis cujas medidas foram propriamente observadas: *temperatura média do ar*, *temperatura máxima do ar*, *temperatura mínima do ar*, *umidade relativa média do ar*, *velocidade do vento* e *radiação*. Todas essas variáveis foram descritas como medidas (*MEASURE*), do tipo número *(NUMERIC*) de formato decimal (*TD NUMERIC*).

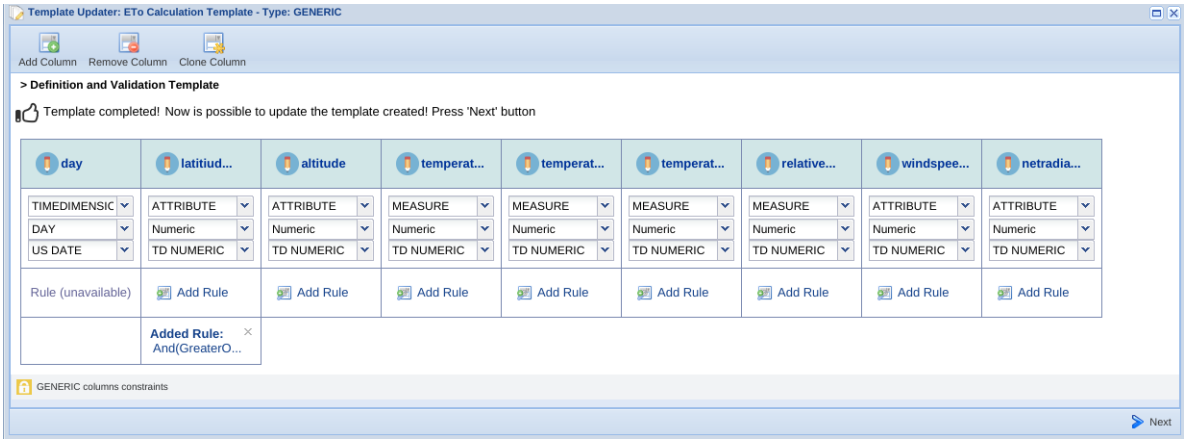
Figura 22: Figura com o criador de Template do TabMan



Fonte: *D4Science Infrasctructure*

Uma outra ferramenta que pode ser explorada é a criação de regras, que podem ser muito úteis para tornar explícito certo conteúdo semântico de uma variável. No caso, é exibida na Figura 24 como ela foi utilizada para limitar a *latitude* à valores semanticamente coerentes com seu conceito geográfico, variando apenas entre -90º e 90º. Existem outros meios de adicionar mais metainformações ao esquema que o conjunto de dados deve se

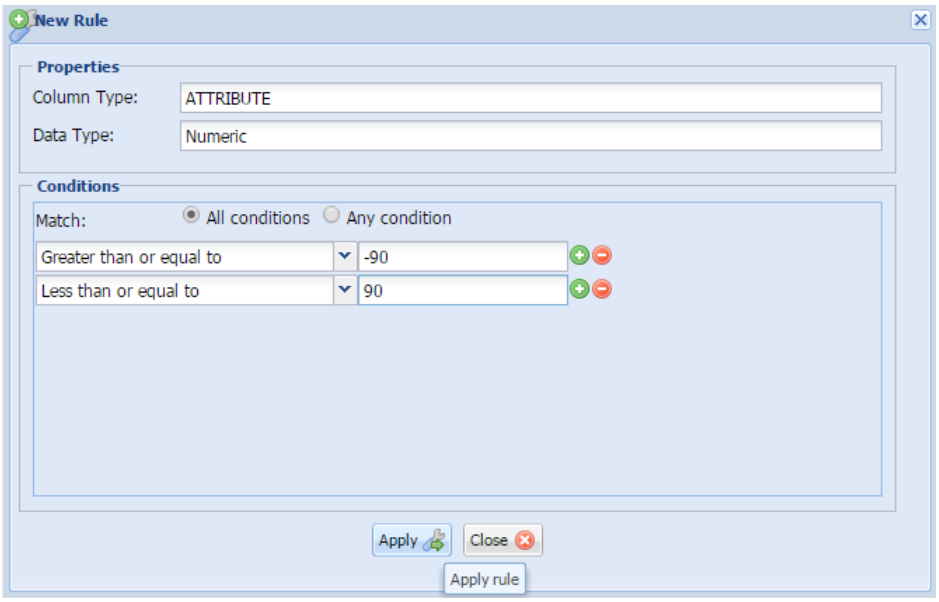
Figura 23: Figura com o criador de Template do TabMan, na tela de configuração das variáveis



Fonte: *D4Science Infrasctructure*

conformar, entre outras coisas, adicionando colunas que derivam seus valores a partir de informações conhecidas pelos campos informados, porém o seu uso não foi considerado adequando para essas situações em questão.

Figura 24: Figura com o criador de Template do TabMan, na tela de criação de regras



Fonte: *D4Science Infrasctructure*

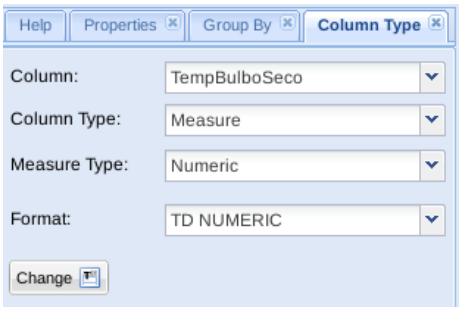
4.1.2.4 Transformações de Dados

Voltando os olhos para a *Situação II* foi necessário realizar um conjunto de transformações semelhantes a situação anterior. Os passos desse processo podem ser descrito

como se segue. Para maior coesão semântica dos *Conjuntos C e D* a primeira dificuldade que tratada foi a relacionada com os ***identificadores***. Como foi explicado a dificuldade semântica reside no fato de se usarem duas codificações distintas para as estações. Devido ao fato de que essa informação referente a origem dos dados não serão mais necessárias e de que o processo de mesclagem dos dados gerará um novo índice pelo qual todas as informações do novo conjunto serão indexadas, além da deleção das variáveis, dispensa-se quaisquer outras providências prévias para coesão final dos dados pelo *TabMan*.

Outro procedimento adotado foi o que resolveu o problema associado ao ***tempo***. Como foi dito os dados do *Conjunto C* estão com periodicidade inferior a diária (3 vezes ao dia), enquanto a resolução temporal do *Conjunto D* é horária. A resolução temporal adotada pelo *template* para homogeneizar a situação foi a diária. Para solucionar isso foi utilizada novamente a ferramenta de agrupamento de variáveis em função de outras. É importante notar que cada variável possui definida na ferramenta a sua própria função de agrumamento que a adéqua à mudança de resolução temporal. Antes disso, foram necessárias as definições dos tipo das variáveis que iriam ser agrupadas como, por exemplo, mostra a Figura 25. Depois dessa preparação, o procedimento de mudança da resolução temporal segue como ilustrado pela Figura 26 e pela Figura 27.

Figura 25: Figura que mostra definição de tipo de uma variável pelo TabMan

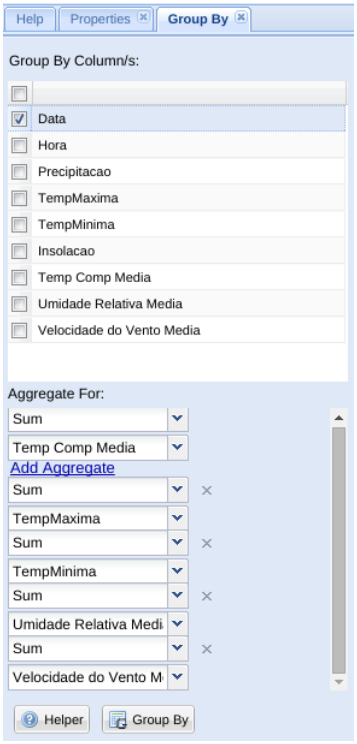


Fonte: *D4Science Infrasctructure*

Outra fonte de diversidade semântica que foi analisada foi a ligada a ***granularidade*** de alguns dados. Com relação aos aspectos que são necessários para atender ao *template* não serão necessários o uso de máximas e mínimas das variáveis de umidade relativa e pressão atmosférica. Nesse caso apenas se seguiu com a remoção dessa variáveis através de ferramenta previamente utilizada. Porém, se fosse interessante aos propósitos de processamento esses dados poderiam ser tratado por meio de transformações diversas como média, por exemplo, a fim de chegarem a uma informação homogênea e de utilidade.

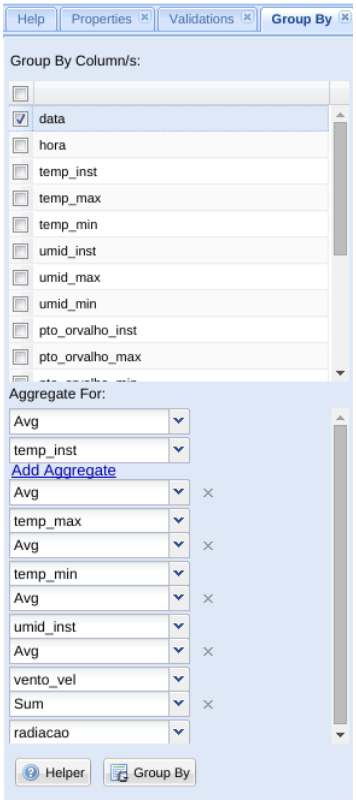
Em seguida, o problema de heterogeneidade semântica relacionado à ***unidade*** de medida foi tratado. Na verdade, essa questão foi resolvida durante e após a mudança da periodicidade dos dados. Ocorreu da seguinte forma. Antes de mudar a resolução temporal

Figura 26: Figura mostra a mudança de resolução temporal do Conjunto C pelo TabMan



Fonte: *D4Science Infrasctructure*

Figura 27: Figura mostra a mudança de resolução temporal do Conjunto D pelo TabMan



Fonte: *D4Science Infrasctructure*

dos dados as medidas de radiação foram multiplicadas por 3600 como manda a regra de conversão. Para isso se fez o uso de uma ferramenta que permite a aplicação de expressões sobre os dados. Depois, durante a agregação dos dados, as medidas foram somadas como aponta a Figura 27. E por fim, após a mudança soma das medidas horárias foi dados foram divididos por 10^6 como define a regra para conversão. Após esse processo essa variável está semanticamente compatível com o *template* de validação.

O último problema de semântica a ser tratado foi o referente ao ***conflito de nomenclatura***. Para solução dessa dificuldade foi utilizada a ferramenta de mudança de rótulos citada anteriormente. Ela adequou todos os campos a nomenclatura adotada nos dois conjuntos pelo *template*.

Uma última preparação para a validação por parte do *template* foi o acréscimo de campos refentes a algumas informações. Primeiramente, explicitando os metadados de localização das estações, nomeadamente a *altitude* e a *latitude*. Em seguida, no caso do *Conjunto C*, acrescentou-se a coluna de *radiação* com valor zero, uma vez que na ausência de dados o algoritmo faz a estimativa dessa variável.

Finalmente, para terminar as outras transformações de tipos necessárias e garantir

uma maior coesão estrutural e semântica foi aplicado o *template* construído para validar os conjuntos de dados da *Situação II*. A aplicação do *template* foi feita com sucesso, fazendo com que os dados estejam preparados para o processamento como mostra, por exemplo, a Figura 28 e Figura 29.

Figura 28: Figura que mostra Conjunto C depois da aplicação do template pelo TabMan

Day	latitude	altitude	temperature	temperaturemax	temperaturemin	relativehumidity	windspeed	radiation
03 Dec 1998	-15.619722	145	26.96	30.1	25.5	89.25	1.366667	0
07 Jun 2002	-15.619722	145	23.9	33.8	15.6	67.75	1.7148	0
29 Sep 2004	-15.619722	145	27.52	32.4	26.3	62	1.9	0
05 Jun 2003	-15.619722	145	22.08	28.2	17.2	90	1	0

Fonte: *D4Science Infrasctructure*

Figura 29: Figura que mostra Conjunto D depois da aplicação do template pelo TabMan

Day	latitude	altitude	temperature	temperaturemax	temperaturemin	relativehumidity	windspeed	radiation
03 May 2009	-15.559295	242	25.25416666666...	26.14583333333...	24.52083333333...	70.20833333333...	0	0
08 Apr 2005	-15.559295	242	28.45833333333...	29.025	27.72916666666...	71.25	1.283333333333...	14.6988
25 Oct 2008	-15.559295	242	28.525	29.02083333333...	27.79583333333...	66.625	1.445833333333...	0
12 Sep 2009	-15.559295	242	22.225	22.77916666666...	21.5125	67.125	0	0

Fonte: *D4Science Infrasctructure*

4.1.2.5 Mescla de Conjuntos de Dados

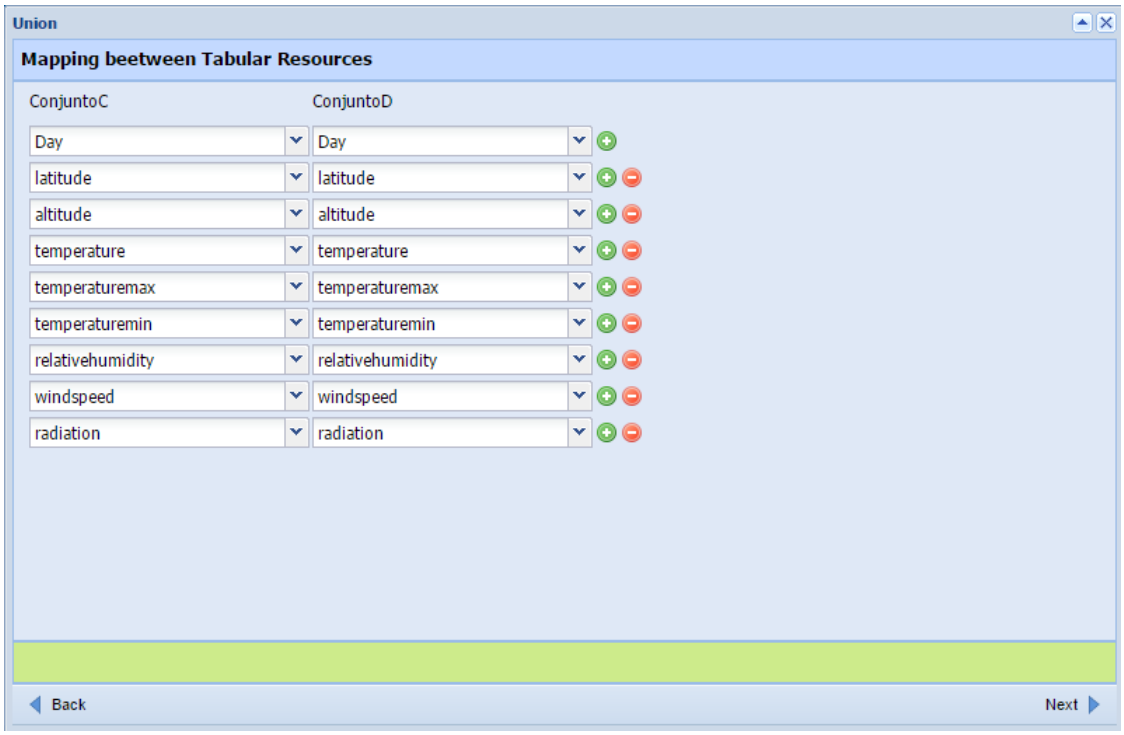
Posteriormente à aplicação do *template* para *Situação II*, os *Conjuntos C e D* foram submetidos ao processo de mesclagem de dados no VRE PGFA-UFMT por meio da ferramenta *StaMan*. Isso fez com que o *Conjunto C* passasse a conter também todos os dados que antes estavam apenas no *Conjunto D*. Isso gerou 13 anos anos de dados sintaticamente e semanticamente coesos distribuídos entre 1998 e 2010. O seu uso se torna mais fácil também pelo compartilhamento que o VRE proporciona. Esse outro procedimento é mostrado na Figura 30.

4.1.2.6 Validação por Processamento Incorporado

Para a segunda situação, foi implementado um algoritmo que realiza o cálculo da estimativa da evapotranspiração de referência pelo método adotado pela *Food and Agriculture Organization* (FAO) tendo como entrada um conjunto de informações micrometeorológicas.

A *Situação II* exigiu uma entrada e saída semelhantes as da situação anterior. Como entrada foi escolhido uma tabela interna do VRE que respeite as definições do *template* feito para a segunda situação. Como saída para o algoritmo foi definido a geração de um recurso tabular dentro do VRE, que traga os dados do recurso de entrada mais

Figura 30: Figura que mostra a mesclagem do Conjunto C e do Conjunto D pelo TabMan



Fonte: *D4Science Infrasctructure*

os relativos a estimativa de evapotranspiração. O processamento para essa situação foi mais complexo, devido a natureza do próprio processamento. De forma geral ele segue as etapas definidas pelo Embrapa para a realização dessas estimativas conforme Conceição (2006). Um diferencial é em relação a interpretação dos dados de entrada. Se os valores da variável de radiação vieram zerados, o algoritmo interpretará que é um sinal para se estimar a radiação por meio de cálculos matemáticos.

Para a implantação se procedeu da mesma maneira que na situação anterior, foi submetido o código-fonte e o pacote de compilados aos administradores da infraestrutura, que são responsáveis pela realização a implantação do algoritmo.

O Conjunto C de dados foi submetido ao processamento programado. O processamento para a *Situação II* trouxe muitos resultados para serem exibidos integralmente neste trabalho, por isso na Tabela 4 optou-se por exibir o ano de junção dos dados: 2004/2005. Nota-se, também, que foi adicionado destaque aos dados de origem do Conjunto D, mas que não cuja fonte de dados e sua semântica original não influíram para o processamento final. Além dessas, outras considerações a respeito de todas as etapas desse processo estão presentes na próxima seção, que elabora uma discussão sobre elas.

Tabela 4: Tabela do resultado do processamento da Situação II.

Mês/Ano	**ETo**
07/2004	86.0115
08/2004	115.592
09/2004	124.687
10/2004	138.161
11/2004	124.675
12/2004	134.281
01/2005	**117.806**
02/2005	**89.1819**
03/2005	**111.296**
04/2005	**101.776**
05/2005	**78.7918**
06/2005	**76.6809**

4.2 Discussão

Algumas considerações a respeito dos diversos procedimentos realizados no Estudo de Caso apresentado serão realizadas nessa seção. O intuito é fazer uma discussão a respeito da heterogeneidade semântica e de que maneira uma HDI como o *D4Science* está preparada para auxiliar os pesquisadores que estão envolvidos com os métodos da Computação Científica nas ciências ambientais.

Este trabalho começou com uma pergunta: uma infraestrutura virtual distribuída como o *D4Science* é capaz de fornecer os meios para mitigar ou superar os problemas de heterogeneidade semânticas que surgem no processo de produção científica em uma ciência que lida com dados ambientais? Na busca de resolver essa questão foi realiza uma pesquisa na literatura a fim de lançar luzes sobre a situação. Como foi visto, a heterogeneidade semântica se apresenta como um fenômeno que está intrinsecamente ligado ao processo de conhecimento humano. Devido a isso, ela apresenta uma de suas faces quando os pesquisadores começam a utilizar as ferramentas computacionais a fim de descreverem seus objetos de pesquisa. Com o advento e popularização da Computação Científica esses problemas se tornaram cada vez mais presentes. Os diversos e frequentes problemas que surgem nesse contexto fez com que a questão de como as técnicas computacionais estavam lidando com esse problema se apresentasse em relevo para essa pesquisa.

Foi realizado apanhado de como as técnicas computacionais estavam se desenvolvendo para um suprir a demanda de terceiro pilar da ciência moderna. Nesse sentido, notou-se que a Computação em Nuvem e as infraestrutura híbridas de dados surgiram com

grande destaque no contexto da Computação Científica. Uma infraestrutura virtual em específico – a *D4Science* – foi escolhida para um estudo de caso devido a sua preocupação em lidar com a heterogeneidade dos dados e fornecer os meios para mitigar a diversidade semântica que proveem deles.

Um roteiro de tratamento dos dados foi proposto a fim de fornecer ao final elementos da capacidade da infraestrutura virtual escolhida lidar com a diversidade semântica. O resultado esperado após esses passos dentro desse Ambiente Virtual de Pesquisa era o de conjuntos de dados que reunissem em si e de forma semanticamente coesa informações provenientes de fontes que inicialmente apresentavam heterogeneidade semântica entre si.

Com relação aos resultados alcançados pode-se ressaltar que as duas situações propostas para o Estudo de Caso puderam dar um visão de geral dos serviços que esse tipo de infraestrutura disponibiliza para o tratamento de inconsistências semânticas. É importante ressaltar que os casos propostos conseguiram abranger vários tipos de heterogeneidade semântica, que foram solucionadas pela aplicação adequadas dos serviços do VRE.

Para a validação dos resultados obtidos pelo tratamento dos dados pelos serviços do VRE foram adotados dois níveis de validação. O primeiro nível consistiu na adoção de um *schema* ou *template* que trazia em si as definições para os conjuntos de dados. Foi observado que os quatro conjuntos de dados de ambas as situações conseguiram passar pela validação de seus respectivos *templates*. Isso significa que os conjuntos de dados que inicialmente eram semanticamente incompatíveis entre si passaram a ser plenamente intercambiáveis, tanto que puderam ser unificados em um único conjunto de dados para cada situação. É interessante destacar as opções adotadas para lidar com cada tipo de heterogeneidade semântica, uma vez que pode ser de utilidade para outros. Na Tabela 5 segue de forma sintetizada a relação entre as categorias de diversidade semântica e as abordagens utilizadas neste trabalho.

Tabela 5: Tabela que relaciona os tipos de heterogeneidade semântica e as abordagens utilizadas para mitigá-los.

Tipo de Heterogeneidade Semântica	**Abordagem**
Tempo	Agregador de variáveis por função
Unidades	Ferramenta de expressões matemáticas
Granularidade	Deleção de detalhes excedentes no conjunto com maior granularidade
Identificadores	Deleção de índices antigos
Nomenclatura	Edição de rótulos das variáveis

O segundo nível de validação foi feito pela realização de processamento sobre os conjuntos de dados já unificados. O teste partiu do seguinte premissa: se os dados são

semanticamente coerentes e não sofrem mais dos problemas de inconsistência semântica que sofriam, eles podem ser usados juntos para um processamento de forma que sua origem seja indiferente. Os resultados apresentados tanto para *Situação I* como para a *Situação II* mostraram que os dados que foram unificados foram capazes de serem processados sem nenhum problema referente a diversidade semântica.

A realização dessas etapas de compatibilização dos dados e validação pelo *template* e pelo processamento mostraram que o VRE oferecido pelo *gCube/D4Science* foi capaz de fornecer os serviços necessários para lidar com os casos comuns de heterogeneidade semântica apontados pela literatura. Essa constatação pesa para a escolha desses ambientes como lugares de trabalho para as comunidades científicas, uma vez que a heterogeneidade semântica dos dados é uma das situação que impõe dificuldades ao compartilhamento e reutilização dos dados coletados. A capacidade de lidar com essas dificuldade faz de infraestruturas, como o *D4Science*, que implementam o conceito de HDI um opção valiosa no contexto da Computação Científica em sua missão de ser um pilar metodológico para as ciências modernas.

Uma consideração importante para se fazer é que, embora as situações propostas tenham abrangido alguns dos tipos de heterogeneidade semântica apontados pela literatura, e os mais comuns no contexto do PGFA, existem outras categorias de diversidade semântica. Dentre elas se pode destacar a de *reutilização de campos*, *codificação* ou *instâncias heterogêneas*. Embora se pudesse simular casos como esses tipos de inconsistência semântica se optou por não tratá-los extensamente neste trabalho pela pouca relevância para o contexto e a facilidade com que eles seriam resolvidos pelas ferramentas do VRE. Esses tipos de diversidade semântica raramente acontecem no contexto de aquisição de séries temporais de dados ambientais. Entretanto, por exemplo, a *reutilização de campo* poderia ser resolvida simplesmente com a utilização de ferramentas de criação de campos e a realocação de informação baseada em filtros que descreveriam a distinção semântica específica do caso. Outro exemplo seria o cenário de *instâncias heterogêneas*, para resolvê-lo o uso de filtros com base em outras informações do registro também seria suficiente.

Além desses tipos de heterogeneidade semântica, a literatura cita a *cardinalidade* ou as correlatas de *conflitos de metadados* e *conflitos estruturais*, que embora não sejam frequentes no contexto de séries temporais poderiam ser interessantes pela complexidade do problema. Essas categorias de diversidade semântica acontecem no nível de modelagem ou em cenários em que várias entidades, com suas séries temporais, podem se associar por relacionamentos. Por exemplo, se fosse necessário fazer com uma série temporal estivesse mais contextualizada com dados de organização, projetos, sensores, tipos de sensores, entre outras informações, como é proposto por Oliveira (2011), seria necessário fazer a união dos vários conjuntos de dados que contêm essas informações no único conjunto de dados da série temporal. Isso ocorre, pois os serviços do VRE não permitem a associação entre

recursos tabulares por meio de relacionamentos. Uma série temporal é tratada como um recurso isolado, e se é necessário acrescentar dados relativos a ela, é preciso que esses dados sejam inseridos nela. Essa abordagem é altamente criticável pela grande replicação de dados e pela confusão semântica que ela gera ao representar vários objetos do mundo real em um único recurso tabular. Um aprimoramento nesse quesito possibilitaria à essa infraestrutura virtual resolver de forma mais eficaz esses casos de diversidade semântica.

É possível também elaborar outras críticas à algumas limitações encontradas na tecnologia disponibilizada pelo *D4Science*. Algumas delas são:

- Uma primeira limitação é a ausência de uma área, ou componente de software, que lide exclusivamente com o processo de homogeneização semântica dos dados. As ferramentas que auxiliaram na solução dos problemas de heterogeneidade semântica estavam inseridas dentro de outros serviços oferecidos pelo VRE.
- Uma segunda limitação é a falta de personalização dos tipos de dados disponíveis para criação de *templates*. Isso mostra um poder limitado de descrição dos metadados que se podem associar aos dados importados ou gerados.
- Outra ausência que foi notada é a incapacidade de se fazer inferências a partir dos metadados dos conjuntos de dados. Apesar de possibilitar a formatação das séries de dados com determinadas descrições de metadados, não existe um meio de se fazer inferências e transformações quem têm como entrada condições que levem em consideração esses metadados.
- Uma fragilidade mais técnica é notada na forma como a API de implementação de processamentos para o StatMan lida com o acesso aos dados. O acesso é realizado de forma direta por meio de declarações em *Structured Query Language* (SQL), comportamento que gera várias questões relativas à segurança e integridade dos dados.

Com base nisso, algumas melhorias podem ser sugeridas a esses ambientes, como: uma maior capacidade de anotação de metadados por parte dos pesquisadores; a possibilidade de adoção de domínios e subdomínios definidos por ontologias; uma maior capacidade personalização dos tipos de dados; uma separação mais clara entre especificações sintáticas e semânticas na definição de *templates*; entre outros pontos. Apesar dessas melhorias, é de reconhecer a boa capacidade de uma infraestrutura virtual como a observada no *framework gCube* e em sua implementação pelo *D4Science*.

4.3 Considerações sobre o Capítulo

Esse capítulo se dedicou a expôr a realização de tarefas teóricas e práticas ligadas aos objetivos dessa pesquisa. Em um primeiro momento se retomou e detalhou os objetivos almejados por esse trabalho: verificar o eficácia do uso de infraestrutura virtuais no tratamento da heterogeneidade semântica. Depois foi proposto um itinerário responsável por realizar essa validação. De forma geral buscou-se (a) realizar um estudo de caso que abordasse caso prático do cotidiano do pesquisador que lida com dados ambientais; (b) elaborar uma discussão a respeito dos resultados obtidos no uso de um VRE como ferramenta para mitigar a diversidade presentes em dados ambientais.

No primeiro ponto, foram realizas atividades ligadas a duas situações dentro do Estudo de Caso. Foram separados conjuntos de dados com heterogeneidade semântica. Os problemas de inconsistência foram localizados e enumerados. Em seguida, passou-se a utilização do ambiente da infraestrutura a fim de resolver os problemas semânticos que os dados possuíam entre si. Adotou-se *schemas* validadores aos quais os conjuntos de dados deveriam se conformar para passar no teste. Depois de diversas transformações os conjuntos foram submetidos a validação pelos *schemas* e eles passaram por essa etapa com sucesso. Posteriormente, para haver mais uma categoria de validação foram realizados processamentos sobre os dados unificados, a fim de verificar se a coesão deles era algo que poderia ser utilizado na prática. Os conjuntos já unificados passaram também nessa verificação.

Uma discussão foi feita visando apontar os pontos positivos e de melhora na infraestrutura *D4Science*, além de comentar como as fases do Estudo de Caso foram realizadas. Nessa questão da discussão foi destacado o Estudo de Caso conseguiu verificar a assertiva inicial para os tipos de heterogeneidade semântica avaliados. No próximo capítulo será realizado um apanhado de todo o trabalho, buscando tecer uma conclusão que conecte tudo o que foi empreendido nessa pesquisa. Além de propôr alguns possíveis caminhos para futuros trabalhos.

***

5 Conclusão

Este trabalho apresentou a validação do uso de serviços de infraestruturas virtuais do tipo *Hybrid Data Infrastructures* como eficazes no tratamento de diversos tipos de heterogeneidade semânticas que surgem no contexto das ciências ambientais. Para isso foi explorado o grande cenário da pesquisa das ciências que lidam com as variáveis ambientais. Dentro do processo de produção científica foram observados dois fenômenos. O primeiro, o advento da *e-Science*, isto é, da adoção da Computação Científica como parte da metodologia científica. E o segundo, a constatação do problema da heterogeneidade semântica, cujas as origens se revelaram enraizadas no próprio processo de conhecimento humano. Esse problema já perceptível na pesquisa tradicional, mas se fez mais fortemente presente com o advento da Computação Científica, justamente pelo poder desta de agregar uma grande quantidade de informação. Com a finalidade de explorar as soluções para mitigar esse problema foi empreendido o esforço deste trabalho de pesquisa. Ele consistiu em algumas etapas que é possível recapitular como segue.

Primeiramente, foi feito um esforço teórico de compreender os principais elementos envolvidos no contexto desse problema. Realizou-se uma pesquisa na literatura que mostrou a origem, os tipos, o contexto e as causas da heterogeneidade semântica. Em seguida foi investigado como esse fenômeno se apresentava dentro do contexto do advento da Computação Científica e as possíveis abordagens para mitigá-lo. Nesse momento as infraestrutura virtuais, que se apresentam como grandes ambientes integrados de pesquisa, se mostraram como alternativas viáveis para se realizar a realização do trabalho de pesquisa como os feitos do contexto da Física Ambiental, e com os meios para lidar com problemas de diversidade semântica dos dados.

No desenvolvimento do trabalho foram adotados materiais que primavam pela sua reusabilidade e escalabilidade, por se tratarem de opções *open source*. Nesse contexto o *framework gCube* e a infraestrutura *D4Science*, que o utiliza, se mostraram ferramentas adequadas para os propósitos do trabalho. Destacou-se, também que esse trabalho é de natureza expositiva, optando por isso pelo desenvolvimento de um estudo de caso, que utiliza uma abordagem qualitativa para cumprir seus objetivos.

Em seguida, um estudo de caso com duas situações comuns para o contexto de pesquisa na Física Ambiental foi realizado. O objetivo principal desse trabalho era demostrar a capacidade de infraestruturas virtuais, como *D4Science*, em lidar com os problemas comuns de heterogeneidade semântica que os pesquisadores enfrentam. Para cumprir com esse objetivo foi desenvolvido um roteiro de atividades. Ele consistiu em coletar fontes de dados que diversas, identificar nelas os tipos de heterogeneidade semântica, utilizar

os serviços que o *D4Science* oferece para solucioná-las, e por validar se as inconsciências semânticas foram eliminadas.

Posteriormente, os resultados deste trabalho foram expostos e discutidos. Eles se mostraram satisfatório pare responder afirmativamente a questão-chave da pesquisa. A infraestrutura virtual escolhida foi capaz de solucionar os problemas de diversidade semânticas presente no ambiente de Computação Científica. Ela entregou conjuntos de dados semanticamente coerentes que puderam ser processados transparentemente por algoritmos. Foram apontadas algumas serviços que poderiam melhorar o tratamento do problema de heterogeneidade semânticas dos dados.

Com a realização de todas essas atividades, se abrangeu o objetivo de verificar se uma infraestrutura virtual de Computação Científica supre as demandas para mitigar heterogeneidade semântica de dados ambientais. Porém, no desenvolvimento dele foram se apresentando possibilidades de expansão da pesquisa desse assunto. Algumas delas se transformaram nas indicações que são elencadas a seguir.

5.1 Contribuições

As principais contribuições feitas pela realização desse trabalho podem ser resumidas como segue:

- O levantamento teórico da problematização de um tema de grande abrangência, mas muitas vezes de difícil observação por sua sutileza dentro do processo de produção científica;
- A adoção de tecnologias *open source* e que permitem a escalabilidade de componentes por meio da implementação de serviços dentro de um API bem definida;
- A definição e configuração dentro da infraestrutura virtual *D4Science* de um VRE que atende as principais necessidades de pesquisa que aparecem no cotidiano de quem lida com dados ambientais, inclusive aqueles semanticamente heterogêneos;
- A identificação dos tipos de heterogeneidade semântica mais comuns nas atividades ligadas ao PGFA;
- A implementação de serviços de processamento, que foram utilizados para verificação dos cenários propostos e incorporados ao VRE;

5.2 Trabalhos Futuros

Do ponto de vista metodológico o Estudo de Caso costuma dar dois resultados: introduzir de forma didática os interessados em um assunto e explorar uma questão que

ainda está em suas fases iniciais de maturação. Neste trabalho ele serviu para as duas coisas. Com relação ao desenvolvimento da questão de pesquisa, ela parece agora ao fim do trabalho muito mais clara. Nesse sentido, levando em conta as limitações encontradas e discutidas, algumas considerações podem ser feitas sobre trabalhos que busquem expandir a investigação científica nesse assunto.

Devido a grande flexibilidade que a tecnologia *gCube* oferece, por se tratar de uma iniciativa *open source* e por ter sido bem arquitetada como *framework*, se pode vislumbrar algumas melhorias de forma mais fácil. Em termos de arquitetura por se tratar de uma infraestrutura orientada a serviço, toda a alteração feita nela pode ser vista com a adição de um serviço aos que ela já oferece. Nesse sentido, alguns serviços adicionais poderiam ser implementados para capacitá-la a lidar com a diversidade semântica de forma mais eficiente.

O *TabMan* é um serviço que oferece uma grande gama de ferramentas que se mostraram suficientes para lidar com os tipos de heterogeneidade semântica encontrados. Porém, um novo serviço voltado exclusivamente para a lidar com a diversidade semântica dos dados seria um instrumento muito mais poderoso para esse finalidade. Ele poderia se chamar *Data Semantic Manager*. Seria um ambiente próprio que trataria desde a importação dos dados, passando por sua descrição semântica utilizando-se de metadados, até a possível transformação automatizada dos dados levando em consideração um determinado *schema* semântico.

Seria interessante que esse serviço inclui-se o suporte à anotação dos dados com metadados por meio de linguagens de ontologias que possui um maior poder descritivo e é personalizável. Um exemplo é o caso da *Web Ontology Language* (OWL).

Outro serviço que ajudaria na mitigação da heterogeneidade semântica dos dados seria um motor de inferência. Ele poderia agir com base nas anotações de metadados de modo a fornecer inferências e transformações dos dados armazenados no VRE em tempo real para outros domínios conforme os requisitos de um *schema* semântico.

O desenvolvimento de um Estudo de Caso que aborde a interação com troca de dados e processamento entre alguns grupos de pesquisas por meio de uma infraestrutura virtual também poderia ser realizado.

Outras atividades de cunho teórico poderiam também ser implementadas. Entre elas, seria interessante um esforço conceitual em desenvolver melhor a quantificação de alguns aspectos da heterogeneidade semântica dentro do contexto da Computação Científica. Essa proposta passaria pela definição de algumas propriedades semânticas dos dados e a adoção de uma métrica para mensurá-las. Isso seria importante, pois possibilitaria o desenvolvimento de técnicas de mensuração de similaridade semântica entre conjuntos de dados.

***

Referências

AMARAL, R. et al. Supporting biodiversity studies with the EUBrazilOpenBio Hybrid Data Infrastructure: EUBRAZILOPENBIO: A DATA INFRASTRUCTURE TO STUDY BIODIVERSITY. *Concurrency and Computation: Practice and Experience*, p. n/a–n/a, mar. 2014. ISSN 15320626. Disponível em: <http://doi.wiley.com/10.1002/cpe.3238>. Citado na página 35.

ARMBRUST, M. et al. A View of Cloud Computing. *Commun. ACM*, v. 53, n. 4, p. 50–58, abr. 2010. ISSN 0001-0782. Disponível em: <http://doi.acm.org/10.1145/1721654.1721672>. Citado 2 vezes nas páginas 24 e 25.

ASLAN, G.; MCLEOD, D. Semantic Heterogeneity Resolution in Federated Databases by Metadata Implantation and Stepwise Evolution. *The VLDB Journal*, v. 8, n. 2, p. 120–132, out. 1999. ISSN 1066-8888. Disponível em: <http://dx.doi.org/10.1007/s007780050077>. Citado na página 43.

ASSANTE, M.; CANDELA, L.; PAGANO, P. An Environment Supporting the Production of Live Research Objects. *Grey Journal (TGJ)*, v. 9, n. 1, 2013. Disponível em: <http://www.opengrey.eu/data/70/01/68/GL14_Assante_et_al_2013_Conference_Preprint.pdf>. Citado 2 vezes nas páginas 30 e 33.

BALAž, A. et al. Development of Grid e-Infrastructure in South-Eastern Europe. *Journal of Grid Computing*, v. 9, n. 2, p. 135–154, jun. 2011. ISSN 1570-7873, 1572-9184. Disponível em: <http://link.springer.com/article/10.1007/s10723-011-9185-0>. Citado na página 27.

BARJAK, F. et al. The Emerging Governance of E-Infrastructure: The Emerging Governance of E-Infrastructure. *Journal of Computer-Mediated Communication*, v. 18, n. 2, p. 1–24, jan. 2013. ISSN 10836101. Disponível em: <http://doi.wiley.com/10.1111/jcc4.12000>. Citado 2 vezes nas páginas 29 e 30.

BERGAMASCHI, S.; CASTANO, S.; VINCINI, M. Semantic Integration of Semistructured and Structured Data Sources. *SIGMOD Rec.*, v. 28, n. 1, p. 54–59, mar. 1999. ISSN 0163-5808. Disponível em: <http://doi.acm.org/10.1145/309844.309897>. Citado 2 vezes nas páginas 40 e 43.

BISHR, Y. Overcoming the semantic and other barriers to GIS interoperability. *International Journal of Geographical Information Science*, v. 12, n. 4, p. 299–314, 1998. ISSN 1365-8816. Disponível em: <http://dx.doi.org/10.1080/136588198241806>. Citado na página 41.

BRIGHT, M. W.; HURSON, A. R.; PAKZAD, S. Automated Resolution of Semantic Heterogeneity in Multidatabases. *ACM Trans. Database Syst.*, v. 19, n. 2, p. 212–253, jun. 1994. ISSN 0362-5915. Disponível em: <http://doi.acm.org/10.1145/176567.176569>. Citado 3 vezes nas páginas 37, 40 e 42.

BUSHHOUSEN, E. Cloud Computing. *Journal of Hospital Librarianship*, v. 11, n. 4, p. 388–392, out. 2011. ISSN 15323269. Disponível em: <http://search.ebscohost.com/login.

aspx?direct=true&db=lih&AN=67054431&lang=pt-br&site=ehost-live&authtype=ip,cookie,uid>. Citado 2 vezes nas páginas 24 e 25.

BUYYA, R.; PANDEY, S.; VECCHIOLA, C. Cloudbus Toolkit for Market-Oriented Cloud Computing. In: JAATUN, M. G.; ZHAO, G.; RONG, C. (Ed.). *Cloud Computing*. Springer Berlin Heidelberg, 2009, (Lecture Notes in Computer Science, 5931). p. 24–44. ISBN 978-3-642-10664-4, 978-3-642-10665-1. Disponível em: <http://link.springer.com/chapter/10.1007/978-3-642-10665-1_4>. Citado na página 24.

BUYYA, R. et al. Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. *Future Generation Computer Systems*, v. 25, n. 6, p. 599–616, jun. 2009. ISSN 0167-739X. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0167739X08001957>. Citado na página 24.

CANDELA, D. C. L. Virtual Research Environments: An Overview and a Research Agenda. *Data Science Journal*, v. 12, p. GRDI75–GRDI81, ago. 2013b. ISSN 1683-1470. Citado 2 vezes nas páginas 33 e 34.

CANDELA, L. et al. Species distribution modeling in the cloud. *Concurrency and Computation: Practice and Experience*, p. n/a–n/a, jul. 2013a. ISSN 1532-0634. Disponível em: <http://onlinelibrary.wiley.com/doi/10.1002/cpe.3030/abstract>. Citado 2 vezes nas páginas 31 e 33.

CANDELA, L. et al. An infrastructure-oriented approach for supporting biodiversity research. *Ecological Informatics*, v. 26, Part 2, n. 0, p. 162 – 172, 2015. ISSN 1574-9541. Information and Decision Support Systems for Agriculture and Environment. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1574954114001022>. Citado na página 44.

CANDELA, L.; CASTELLI, D.; PAGANO, P. Managing big data through hybrid data infrastructures. *ERCIM News*, v. 2012, n. 89, 2012. Disponível em: <http://dblp.uni-trier.de/db/journals/ercim/ercim2012.html#CandelaCP12>. Citado na página 31.

CERI, S.; WIDOM, J. Managing Semantic Heterogeneity with Production Rules and Persistent Queues. In: . Dublin, Ireland: [s.n.], 1993. Disponível em: <http://ilpubs.stanford.edu:8090/24/>. Citado 4 vezes nas páginas 37, 38, 42 e 56.

CHEPTSOV, A. et al. e-Infrastructure for Remote Instrumentation. *Computer Standards & Interfaces*, v. 34, n. 6, p. 476–484, nov. 2012. ISSN 09205489. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S0920548911001140>. Citado na página 29.

COLOMB, R. M. Impact of Semantic Heterogeneity on Federating Databases. *The Computer Journal*, v. 40, n. 5, p. 235–244, jan. 1997. ISSN 0010-4620, 1460-2067. Disponível em: <http://comjnl.oxfordjournals.org/content/40/5/235>. Citado 2 vezes nas páginas 40 e 42.

CONCEIçãO, M. A. F. *Roteiro de cálculo da evapotranspiração de referência pelo método de Penman-Monteith-FAO*. [S.l.], 2006. In Portuguese, 8 pages. Citado 2 vezes nas páginas 72 e 77.

COSSU, R. et al. A roadmap for a dedicated Earth Science Grid platform. *Earth Science Informatics*, v. 3, n. 3, p. 135–148, set. 2010. ISSN 1865-0473, 1865-0481. Disponível em: <http://link.springer.com/10.1007/s12145-010-0045-4>. Citado na página 28.

D4SCIENCE INFRASTRUCTURE. *D4Science Infrastructure: About Us*. 2015a. Disponível em: <https://www.d4science.org/about-us>. Acesso em: 12 mai. 2015. Citado na página 34.

D4SCIENCE INFRASTRUCTURE. *D4Science Infrastructure: Infrastructure Capacity*. 2015b. Disponível em: <https://www.d4science.org/infrastructure-capacity>. Acesso em: 12 mai. 2015. Citado 2 vezes nas páginas 34 e 47.

D4SCIENCE INFRASTRUCTURE. *D4Science Infrastructure: Exploitation Models*. 2015c. Disponível em: <https://www.d4science.org/exploitation-models>. Acesso em: 13 mai. 2015. Citado 2 vezes nas páginas 35 e 46.

DRAKE, N. Cloud computing beckons scientists. *Nature*, v. 509, p. 543–544, 2014. Disponível em: <http://books.google.com/books?hl=en&lr=&id=KqJVKs6H13AC&oi=fnd&pg=PA13&dq=%22vehicles.+%E2%80%9CIt%E2%80%99s+pretty%22+%22thicker,+multi-year+ice.+(In+line+with%22+%22science+vessel+is+that+it+will+not%22+%22from+its%22+%22on,+through+the+Panama+Canal.%22+%22costs+aside,+the+cloud+will+probably%22+&ots=yhphrSGgXK&sig=Thx7wdXcHYRueEkpdmgeqPpNmM0>. Citado na página 26.

ERIKSSON, O.; GOLDKUHL, G. Preconditions for public sector e-infrastructure development. *Information and Organization*, v. 23, n. 3, p. 149–176, jul. 2013. ISSN 14717727. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S147177271300016X>. Citado na página 28.

FANG, D.; HAMMER, J.; MCLEOD, D. The identification and resolution of semantic heterogeneity in multidatabase systems. *Multidatabase Systems: An Advanced Solution for Global Information Sharing*, p. 52–60, 1994. Disponível em: <http://dbpubs.stanford.edu/pub/showDoc.Fulltext?lang=en&doc=1991-1&format=ps&compression=&name=1991-1.ps>. Citado 2 vezes nas páginas 41 e 42.

FOSTER, I. et al. Cloud Computing and Grid Computing 360-Degree Compared. In: *Grid Computing Environments Workshop, 2008. GCE '08*. [S.l.: s.n.], 2008. p. 1–10. Citado na página 24.

FOX, A. Cloud Computing–What's in It for Me as a Scientist? *Science*, v. 331, n. 6016, p. 406–407, jan. 2011. ISSN 0036-8075, 1095-9203. Disponível em: <http://www.sciencemag.org/cgi/doi/10.1126/science.1198981>. Citado na página 25.

GCUBE CONSORTIUM. *What it is gCube?* 2015. Disponível em: <https://www.gcube-system.org/what-it-is>. Acesso em: 16 mar. 2015. Citado 4 vezes nas páginas 31, 32, 33 e 45.

GCUBE CONSORTIUM. *How-to Implement Algorithms for the Statistical Manager*. 2015b. Disponível em: <http://gcube.wiki.gcube-system.org/gcube/index.php/How-to_Implement_Algorithms_for_the_Statistical_Manager>. Acesso em: 29 jun. 2015. Citado na página 66.

GRACIA, J.; MENA, E. Semantic Heterogeneity Issues on the Web. *IEEE Internet Computing*, v. 16, n. 5, p. 60–67, set. 2012. ISSN 1089-7801. Citado na página 41.

HAKIMPOUR, F.; GEPPERT, A. Resolving Semantic Heterogeneity in Schema Integration. In: *Proceedings of the International Conference on Formal Ontology in Information Systems - Volume 2001*. New York, NY, USA: ACM, 2001. (FOIS '01), p. 297–308. ISBN 1-58113-377-4. Disponível em: <http://doi.acm.org/10.1145/505168.505196>. Citado 2 vezes nas páginas 39 e 43.

HAKIMPOUR, F.; TIMPF, S. Using Ontologies for Resolution of Semantic Heterogeneity in GIS. In: *in GIS. Proceedings 4th AGILE Conference on Geographic Information Science*. [S.l.: s.n.], 2001. p. 385–395. Citado 3 vezes nas páginas 37, 41 e 43.

HAMMER, J.; MCLEOD, D. AN APPROACH TO RESOLVING SEMANTIC HETEROGENEITY IN A FEDERATION OF AUTONOMOUS, HETEROGENEOUS DATABASE SYSTEMS. *International Journal of Cooperative Information Systems*, v. 02, n. 01, p. 51–83, mar. 1993. ISSN 0218-8430. Disponível em: <http://www.worldscientific.com/doi/abs/10.1142/S0218215793000046>. Citado 2 vezes nas páginas 41 e 42.

HAN, L. et al. FireGrid: An e-infrastructure for next-generation emergency response support. *Journal of Parallel and Distributed Computing*, v. 70, n. 11, p. 1128–1141, nov. 2010. ISSN 0743-7315. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0743731510001164>. Citado na página 27.

HELLWEG, H. et al. Treatment of Semantic Heterogeneity in Information Retrieval. *arXiv:1102.3866 [cs]*, fev. 2011. Disponível em: <http://arxiv.org/abs/1102.3866>. Citado na página 37.

HEY, T.; TANSLEY, S.; TOLLE, K. M. Jim gray on escience: a transformed scientific method. In: HEY, T.; TANSLEY, S.; TOLLE, K. M. (Ed.). *The Fourth Paradigm*. Microsoft Research, 2009. ISBN 978-0982544204. Disponível em: <http://dblp.uni-trier.de/db/books/collections/4paradigm2009.html#HeyTT09>. Citado na página 29.

HEY, T.; TREFETHEN, A. E. Cyberinfrastructure for e. *Science*, v. 308, n. 5723, p. 817–821, jun. 2005. ISSN 0036-8075, 1095-9203. Disponível em: <http://www.sciencemag.org/content/308/5723/817>. Citado 2 vezes nas páginas 26 e 27.

HOLTIES, H.; RENTING, A.; GRANGE, Y. The LOFAR long-term archive: e-infrastructure on petabyte scale. In: RADZIWILL, N. M.; CHIOZZI, G. (Ed.). [s.n.], 2012. p. 845117. Disponível em: <http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.927147>. Citado na página 28.

HORSBURGH, J. S. et al. Managing a community shared vocabulary for hydrologic observations. *Environmental Modelling & Software*, v. 52, p. 62–73, fev. 2014. ISSN 1364-8152. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1364815213002557>. Citado 3 vezes nas páginas 38, 42 e 43.

HULL, R. Managing Semantic Heterogeneity in Databases: A Theoretical Prospective. In: *Proceedings of the Sixteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles*

of Database Systems. New York, NY, USA: ACM, 1997. (PODS '97), p. 51–61. ISBN 0-89791-910-6. Disponível em: <http://doi.acm.org/10.1145/263661.263668>. Citado 2 vezes nas páginas 37 e 39.

IBM CORPORATION. *IBM SPSS Modeler CRISP-DM Guide*. [S.l.], 2011. Disponível em: <ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/14.2/en/CRISP_DM.pdf>. Citado 2 vezes nas páginas 48 e 49.

KASHYAP, V.; SHETH, A. Semantic Heterogeneity in Global Information Systems: The Role of Metadata, Context and Ontologies. In: PAPAZOGLOU, M. P.; SCHLAGETER, G. (Ed.). *Cooperative Information Systems*. San Diego: Academic Press, 1998. p. 139–178. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.43.3117>. Citado 2 vezes nas páginas 40 e 42.

KNIGHT, N.; POULSON, J. Scientific Computing. *XRDS*, v. 19, n. 3, p. 7–7, mar. 2013. ISSN 1528-4972. Disponível em: <http://doi.acm.org.ez52.periodicos.capes.gov.br/10.1145/2425676.2425679>. Citado na página 22.

LAKATOS, E. M.; MARCONI, M. de A. *Metodologia Científica*. São Paulo: Atlas S.A, 1986. 231 p. Citado na página 47.

LECCA, G. et al. Grid computing technology for hydrological applications. *Journal of Hydrology*, v. 403, n. 1-2, p. 186–199, jun. 2011. ISSN 00221694. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S0022169411002435>. Citado na página 28.

MARCONI, M. de A.; LAKATOS, E. M. *Fundamentos da Metodologia Científica*. 5. ed. São Paulo: Atlas S.A, 2003. 305 p. Citado na página 47.

MELL, P.; GRANCE, T. The NIST definition of cloud computing. 2011. Disponível em: <http://faculty.winthrop.edu/domanm/csci411/Handouts/NIST.pdf>. Citado na página 25.

MITRA, P.; WIEDERHOLD, G. Resolving terminological heterogeneity in ontologies. In: *Proceedings of the ECAI workshop on Ontologies and Semantic Interoperability*. [s.n.], 2002. Disponível em: <http://iwayan.info/Research/Ontology/Papers_Research/OntoMapping/Mitra02_ResolvingTerminologicalOntology.pdf>. Citado 2 vezes nas páginas 37 e 39.

OLIVEIRA, A. G. de. *Arquitetura de Dados Para Gerenciamento de Informações*. 86 p. Dissertação (Mestrado em Física Ambiental) — Instituto de Física, Universidade Federal de Mato Grosso, Cuiabá, 2011. Citado na página 80.

OSTERMANN, S. et al. A Performance Analysis of EC2 Cloud Computing Services for Scientific Computing. In: AVRESKY, D. R. et al. (Ed.). *Cloud Computing*. Springer Berlin Heidelberg, 2010, (Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, 34). p. 115–131. ISBN 978-3-642-12635-2, 978-3-642-12636-9. Disponível em: <http://link.springer.com/chapter/10.1007/978-3-642-12636-9_9>. Citado na página 25.

PARASHAR, M.; THIRUVATHUKAL, G. K. Cloud Computing. *Computing in Science & Engineering*, v. 15, n. 4, p. 0008–9, 2013. Disponível em: <http://www.computer.org/csdl/mags/cs/2013/04/mcs2013040008.html>. Citado na página 26.

PEREZ, F.; GRANGER, B. E.; HUNTER, J. D. Python: an ecosystem for scientific computing. *Computing in Science & Engineering*, v. 13, n. 2, p. 13–21, 2011. Disponível em: <http://scitation.aip.org/content/aip/journal/cise/13/2/10.1109/MCSE.2010.119>. Citado na página 22.

PRODAN, R.; SPERK, M. Scientific computing with Google App Engine. *Future Generation Computer Systems*, v. 29, n. 7, p. 1851–1859, set. 2013. ISSN 0167739X. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S0167739X13000022>. Citado na página 23.

RUTTIMANN, J. 2020 computing: Milestones in scientific computing. *Nature*, v. 440, n. 7083, p. 399–405, mar. 2006. ISSN 0028-0836, 1476-4679. Disponível em: <http://www-nature-com.ez52.periodicos.capes.gov.br/nature/journal/v440/n7083/full/440399a.html>. Citado na página 21.

SAGER, S.; MOMBAUR, K.; FUNKE, J. Scientific computing for the cognitive sciences. *Journal of Computational Science*, v. 4, n. 4, p. 242–244, jul. 2013. ISSN 18777503. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S187775031200124X>. Citado na página 22.

SEIDEL, E. L. Metadata Management in Scientific Computing. *arXiv:1203.4135 [cs]*, mar. 2012. ArXiv: 1203.4135. Disponível em: <http://arxiv.org/abs/1203.4135>. Citado na página 23.

SHETH, A. P.; LARSON, J. A. Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases. *ACM Comput. Surv.*, v. 22, n. 3, p. 183–236, set. 1990. ISSN 0360-0300. Disponível em: <http://doi.acm.org/10.1145/96602.96604>. Citado na página 36.

SHVAIKO, P. et al. Web Explanations for Semantic Heterogeneity Discovery. In: GóMEZ-PéREZ, A.; EUZENAT, J. (Ed.). *The Semantic Web: Research and Applications*. Springer Berlin Heidelberg, 2005, (Lecture Notes in Computer Science, 3532). p. 303–317. ISBN 978-3-540-26124-7, 978-3-540-31547-6. Disponível em: <http://link.springer.com/chapter/10.1007/11431053_21>. Citado 2 vezes nas páginas 41 e 43.

SINGH, M. A framework for data modeling and querying dataspace systems. *Seventh International Conference on Data Mining and Warehousing (ICDMW-2013)*, p. 17–25, 2013. Disponível em: <http://searchdl.org/index.php/book_series/downloadPDF/880>. Citado 3 vezes nas páginas 38, 41 e 43.

SZALAY, A. Extreme data-intensive scientific computing. *Computing in Science & Engineering*, v. 13, n. 6, p. 34–41, 2011. Disponível em: <http://scitation.aip.org/content/aip/journal/cise/13/6/10.1109/MCSE.2011.74>. Citado 2 vezes nas páginas 22 e 23.

VENTRONE, V.; HEILER, S. Semantic Heterogeneity As a Result of Domain Evolution. *SIGMOD Rec.*, v. 20, n. 4, p. 16–20, dez. 1991. ISSN 0163-5808. Disponível em: <http://doi.acm.org/10.1145/141356.141359>. Citado 6 vezes nas páginas 37, 38, 39, 40, 42 e 56.

WALKER, D. W. et al. Selected papers from the 2010 e-Science All Hands Meeting. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering*

Sciences, v. 369, n. 1949, p. 3251–3253, ago. 2011. ISSN 1364-503X, 1471-2962. Disponível em: <http://rsta.royalsocietypublishing.org/cgi/doi/10.1098/rsta.2011.0147>. Citado na página 29.

WIKIMEDIA COMMONS. *CRISP-DM Process Diagram*. 2015. Disponível em: <https://commons.wikimedia.org/wiki/File:CRISP-DM_Process_Diagram.png>. Acesso em: 16 out. 2015. Citado na página 49.

WORBOYS, M. F.; DEEN, S. M. Semantic Heterogeneity in Distributed Geographic Databases. *SIGMOD Rec.*, v. 20, n. 4, p. 30–34, dez. 1991. ISSN 0163-5808. Disponível em: <http://doi.acm.org/10.1145/141356.141366>. Citado na página 38.

Apêndices

APÊNDICE A – Trechos do Algoritmo de Precipitação Anual

```
1  package org.gcube.dataanalysis.precipitations;
2
3  (...)
4
5  public class AnnualPrecipitation extends StandardLocalExternalAlgorithm {
6    // Class Attributes
7    String outputtablename;
8    String outputtable;
9
10 (...)
11
12   @Override
13   protected void process() throws Exception {
14     // Recovering data
15     config.setParam("DatabaseDriver", "org.postgresql.Driver");
16     SessionFactory dbconnection = DatabaseUtils.initDBSession(config);
17     try {
18       String tablename = getInputParameter("PrecTable");
19       String columnnames = getInputParameter("PrecColumns");
20       outputtablename = getInputParameter("OutputTableName");
21       outputtable = getInputParameter("OutputTable");
22       String[] columnlist = columnnames.split(AlgorithmConfiguration.
             getListSeparator());
23       List<Object> dataList = DatabaseFactory.executeSQLQuery("select " +
             columnlist[0] + " from " + tablename + " order by " + columnlist[0] +
             " asc", dbconnection);
24       List<Object> precList = DatabaseFactory.executeSQLQuery("select " +
             columnlist[1] + " from " + tablename + " order by " + columnlist[0] +
             " asc", dbconnection);
25       // Business Logic
26       AnalysisLogger.getLogger().info("Creating output table [" + "create table
             " + outputtable + " (year integer, value real)]");
27       DatabaseFactory.executeSQLUpdate("create table " + outputtable + " (year
             integer, value real)", dbconnection);
28       Double averageValue = new Double(0);
29       for (int i = 0; i < dataList.size(); i++) { // for each line
30         if (i == 0) { // first iteration
31           averageValue = averageValue + Double.parseDouble(String.valueOf(
                 precList.get(i)));
32           if (dataList.size() == 1) { // if first iteration is also
33                         // the last
34             Date currentDate = anyStringToDate(String.valueOf(dataList.get(i)));
35             Calendar currentDateCal = Calendar.getInstance();
36             currentDateCal.setTime(currentDate);
37             int yearCurrentDate = currentDateCal.get(Calendar.YEAR);
38             AnalysisLogger.getLogger().info("Inserting into table " + "insert
                   into " + outputtable + " (year,value) values (" +
                   yearCurrentDate + "," + averageValue + ")");
```

```
39              DatabaseFactory.executeSQLUpdate("insert into " + outputtable + " (
                    year,value) values (" + yearCurrentDate + "," + averageValue + "
                    )", dbconnection);
40              averageValue = new Double(0);
41            }
42          }
43          if (i > 0) { // other iterations
44            Date currentDate = anyStringToDate(String.valueOf(dataList.get(i)));
45            Date lastDate = anyStringToDate(String.valueOf(dataList.get(i - 1)));
46            Calendar currentDateCal = Calendar.getInstance();
47            currentDateCal.setTime(currentDate);
48            int yearCurrentDate = currentDateCal.get(Calendar.YEAR);
49            Calendar lastDateCal = Calendar.getInstance();
50            lastDateCal.setTime(lastDate);
51            int yearLastDate = lastDateCal.get(Calendar.YEAR);
52            if (yearCurrentDate > yearLastDate) {
53              // inserting average annual
54              AnalysisLogger.getLogger().info("Inserting into table " + "insert
                    into " + outputtable + " (year,value) values (" + yearLastDate +
                     "," + averageValue + ")");
55              DatabaseFactory.executeSQLUpdate("insert into " + outputtable + " (
                    year,value) values (" + yearLastDate + "," + averageValue + ")",
                     dbconnection);
56              averageValue = new Double(0);
57              averageValue = new Double(String.valueOf(precList.get(i)));
58            } else {
59              averageValue = averageValue + Double.parseDouble(String.valueOf(
                    precList.get(i)));
60            }
61          }
62          if (i != 0 && i == dataList.size() - 1) { // last iteration
63            Date currentDate = anyStringToDate(String.valueOf(dataList.get(i)));
64            Calendar currentDateCal = Calendar.getInstance();
65            currentDateCal.setTime(currentDate);
66            int yearCurrentDate = currentDateCal.get(Calendar.YEAR);
67            AnalysisLogger.getLogger().info("Inserting into table " + "insert into
                     " + outputtable + " (year,value) values (" + yearCurrentDate + ",
                    " + averageValue + ")");
68            DatabaseFactory.executeSQLUpdate("insert into " + outputtable + " (
                    year,value) values (" + yearCurrentDate + "," + averageValue + ")"
                    , dbconnection);
69            averageValue = new Double(0);
70          }
71        }
72      } catch (Exception e) {
73        AnalysisLogger.getLogger().error(e.getMessage());
74        throw e;
75      } finally {
76        DatabaseUtils.closeDBConnection(dbconnection);
77      }
78    }
79
80  (...)
81
82  }
```

APÊNDICE B – Trechos do Algoritmo para Cálculo da Estimativa de Evapotranspiração de Referência pelo método de Penman–Monteith–FAO

```
1  package org.gcube.dataanalysis.etopenmanmonteithfao;
2
3  (...)
4  /**
5   * This is an algorithm that returns the the calculation of The Penman-Monteith
        Evapotranspiration Estimation by FAO-56 Method. The input is a general
        tabular resource with eight columns (date,
6   * latitude, altitude, temperature, max temperature, min temperature relative
        humidity, wind speed, and radiation).
7   */
8  public class EtoPenmanMonteithFao extends StandardLocalExternalAlgorithm {
9    // Class Attributes
10   String outputtablename;
11   String outputtable;
12
13 (...)
14
15   @Override
16   protected void process() throws Exception {
17     // Recovering data
18     config.setParam("DatabaseDriver", "org.postgresql.Driver");
19     SessionFactory dbconnection = DatabaseUtils.initDBSession(config);
20     try {
21       // Getting context variables
22       String tablename = getInputParameter("MetTable");
23       String columnnames = getInputParameter("MetColumns");
24       outputtablename = getInputParameter("OutputTableName");
25       outputtable = getInputParameter("OutputTable");
26       // Select variables from tables
27       String[] columnlist = columnnames.split(AlgorithmConfiguration.
             getListSeparator());
28       List<Object> dayList = DatabaseFactory.executeSQLQuery("select " +
             columnlist[0] + " from " + tablename + " order by " + columnlist[0] +
             " asc", dbconnection);
29       List<Object> latitudeList = DatabaseFactory.executeSQLQuery("select " +
             columnlist[1] + " from " + tablename + " order by " + columnlist[0] +
             " asc", dbconnection);
30       List<Object> altitudeList = DatabaseFactory.executeSQLQuery("select " +
             columnlist[2] + " from " + tablename + " order by " + columnlist[0] +
```

```
            " asc", dbconnection);
31          List<Object> tempatureList = DatabaseFactory.executeSQLQuery("select " +
                columnlist[3] + " from " + tablename + " order by " + columnlist[0] +
                " asc", dbconnection);
32          List<Object> maxTempatureList = DatabaseFactory.executeSQLQuery("select "
                + columnlist[4] + " from " + tablename + " order by " + columnlist[0]
                + " asc", dbconnection);
33          List<Object> minTempatureList = DatabaseFactory.executeSQLQuery("select "
                + columnlist[5] + " from " + tablename + " order by " + columnlist[0]
                + " asc", dbconnection);
34          List<Object> relativeHumidityList = DatabaseFactory.executeSQLQuery("
                select " + columnlist[6] + " from " + tablename + " order by " +
                columnlist[0] + " asc", dbconnection);
35          List<Object> windSpeedList = DatabaseFactory.executeSQLQuery("select " +
                columnlist[7] + " from " + tablename + " order by " + columnlist[0] +
                " asc", dbconnection);
36          List<Object> radiationList = DatabaseFactory.executeSQLQuery("select " +
                columnlist[8] + " from " + tablename + " order by " + columnlist[0] +
                " asc", dbconnection);
37          // creating output table
38          AnalysisLogger.getLogger().info("Creating output table [" + "create table
                " + outputtable + " (day date, etopmf real)]");
39          DatabaseFactory.executeSQLUpdate("create table " + outputtable + " (day
                date, etopmf real)", dbconnection);
40          /** Business Logic **/
41          // Estimation of reference evapotranspiration
42          /** Variables Declaration **/
43          Date day;
44          Double etopmf; // It is reference evapotranspiration (mm)
45          Double rs; // It is the balance of daily radiation
46          Double lambda; // is the psychrometric coefficient
47          Double t; // It is the average air temperature
48          Double maxT; // It is the max air temperature
49          Double minT; // It is the min air temperature
50          Double u2; // It is the wind speed at 2m height
51          Double z; // It is the place altitude (m)
52          Double patm; // It is atmospheric pressure (kPa)
53          Double ru; // It is relative humidity (%)
54          Double lat; // It is latitude (decimal coordinates)
55          /** Variable Assignments - Variables that is the same for whole Data Set *
                */
56          // altitude
57          z = Double.parseDouble(String.valueOf(altitudeList.get(0)));
58          // latitude
59          lat = Double.parseDouble(String.valueOf(latitudeList.get(0)));
60          // etopmf
61          etopmf = 0.0D;
62          /** Critic on variables values **/
63          Boolean isLineWithNullValues = false;
64          if (z == null) {
65            isLineWithNullValues = true;
66          }
67          if (lat == null) {
68            isLineWithNullValues = true;
69          }
70          if (!isLineWithNullValues) {
71            /** Constants Calculation - Variables that changes everyday **/
```

```
72          // patm
73          patm = (Double) (101.3 * Math.pow(((293 - 0.0065 * z) / (293)), 5.26));
74          // lambda
75          lambda = (Double) (0.665 * Math.pow(10, -3) * patm);
76          // for each line
77          for (int i = 0; i < dayList.size(); i++) {
78            /** Variable Assignments **/
79            // temperature
80            t = Double.parseDouble(String.valueOf(tempatureList.get(i)));
81            // max temperature
82            maxT = Double.parseDouble(String.valueOf(maxTempatureList.get(i)));
83            // temperature
84            minT = Double.parseDouble(String.valueOf(minTempatureList.get(i)));
85            // relative humidity
86            ru = Double.parseDouble(String.valueOf(relativeHumidityList.get(i)));
87            // wind speed
88            u2 = Double.parseDouble(String.valueOf(windSpeedList.get(i)));
89            // rn
90            rs = Double.parseDouble(String.valueOf(radiationList.get(i)));
91            // day
92            day = anyStringToDate(String.valueOf(dayList.get(i)));
93            /** Critic on variables values **/
94            if (t == null) {
95              isLineWithNullValues = true;
96            }
97            if (maxT == null) {
98              isLineWithNullValues = true;
99            }
100           if (minT == null) {
101             isLineWithNullValues = true;
102           }
103           if (maxT < minT) {
104             isLineWithNullValues = true;
105           }
106           if (ru == null) {
107             isLineWithNullValues = true;
108           }
109           if (u2 == null) {
110             isLineWithNullValues = true;
111           }
112           if (day == null) {
113             isLineWithNullValues = true;
114           }
115           if (!isLineWithNullValues) {
116             // etopmf = dailyEtoCalculation(day, lat, z, t, maxT, minT, ru, u2,
                  rs, lambda);
117             if (i == 0) { // first iteration
118               etopmf = dailyEtoCalculation(day, lat, z, t, maxT, minT, ru, u2,
                    rs, lambda);
119               if (dayList.size() == 1) { // if first iteration is also
120                                          // the last
121                 // inserting etopmf
122                 SimpleDateFormat sdf = new SimpleDateFormat("yyyy/MM");
123                 String dateStr = sdf.format(anyStringToDate(String.valueOf(
                      dayList.get(i))));
124                 AnalysisLogger.getLogger().info("Inserting into table "
125                     + "insert into "
```

```
126                     + outputtable
127                     + " (day, etopmf)  values (to_date('"
128                     + dateStr
129                     + "', 'yyyy/MM'),"
130                     + etopmf
131                     + ")");
132                   DatabaseFactory.executeSQLUpdate("insert into " + outputtable +
                          " (day,etopmf) values (to_date('" + dateStr + "', 'yyyy/MM')
                          ," + etopmf + ")", dbconnection);
133                   etopmf = new Double(0);
134                 }
135               }
136               if (i > 0) { // other iterations
137                 Date currentDate = anyStringToDate(String.valueOf(dayList.get(i)))
                        ;
138                 Date lastDate = anyStringToDate(String.valueOf(dayList.get(i - 1))
                        );
139                 Calendar currentDateCal = Calendar.getInstance();
140                 currentDateCal.setTime(currentDate);
141                 int monthCurrentDate = currentDateCal.get(Calendar.MONTH);
142                 int yearCurrentDate = currentDateCal.get(Calendar.YEAR);
143                 Calendar lastDateCal = Calendar.getInstance();
144                 lastDateCal.setTime(lastDate);
145                 int monthLastDate = lastDateCal.get(Calendar.MONTH);
146                 int yearLastDate = lastDateCal.get(Calendar.YEAR);
147                 if (monthCurrentDate > monthLastDate || yearCurrentDate >
                        yearLastDate) {
148                   SimpleDateFormat sdf = new SimpleDateFormat("yyyy/MM");
149                   String dateStr = sdf.format(lastDate);
150                   AnalysisLogger.getLogger().info("Inserting into table "
151                       + "insert into "
152                       + outputtable
153                       + " (day, etopmf)  values (to_date('"
154                       + dateStr
155                       + "', 'yyyy/MM'),"
156                       + etopmf
157                       + ")");
158                   DatabaseFactory.executeSQLUpdate("insert into " + outputtable +
                          " (day,etopmf) values (to_date('" + dateStr + "', 'yyyy/MM')
                          ," + etopmf + ")", dbconnection);
159                   etopmf = new Double(0);
160                   etopmf = dailyEtoCalculation(day, lat, z, t, maxT, minT, ru, u2,
                           rs, lambda);
161                 } else {
162                   etopmf = etopmf + dailyEtoCalculation(day, lat, z, t, maxT, minT
                          , ru, u2, rs, lambda);
163                 }
164               }
165               if (i != 0 && i == dayList.size() - 1) { // last iteration
166                 // inserting etopmf
167                 SimpleDateFormat sdf = new SimpleDateFormat("yyyy/MM");
168                 String dateStr = sdf.format(anyStringToDate(String.valueOf(dayList
                        .get(i))));
169                 AnalysisLogger.getLogger().info("Inserting into table " + "insert
                        into " + outputtable + " (day, etopmf)  values (to_date('" +
                        dateStr + "', 'yyyy/MM'),"  + etopmf + ")");
170                 DatabaseFactory.executeSQLUpdate("insert into " + outputtable + "
```

```
                    (day,etopmf) values (to_date('" + dateStr + "', 'yyyy/MM'),(" +
                     etopmf + ")", dbconnection);
171               }
172             }
173             // Set null to all internal variables
174             t = null;
175             ru = null;
176             rs = null;
177             isLineWithNullValues = false;
178           }
179         } else {
180           throw new Exception("No altitude or latitude value defined for this
                tabular resource.");
181         }
182       } catch (Exception e) {
183         AnalysisLogger.getLogger().error(e.getMessage());
184         throw e;
185       } finally {
186         DatabaseUtils.closeDBConnection(dbconnection);
187       }
188     }
189
190   (...)
191
192     /**
193      * An auxiliary method that return the ETo estimation of the day by Penman-
            Monteith-FAO method.
194      *
195      * @param day
196      * @param lat
197      * @param z
198      * @param t
199      * @param maxT
200      * @param minT
201      * @param ru
202      * @param u2
203      * @param rn
204      * @param lambda
205      * @return
206      */
207     public static Double dailyEtoCalculation(Date day, Double lat, Double z,
          Double t, Double maxT, Double minT, Double ru, Double u2, Double rs,
          Double lambda) {
208       Double dailyetopmf = new Double(0);
209       /** Calculation **/
210       // g - It is the total daily flux of heat in the soil
211       Double g = 0D;
212       // delta - It is the slope of the vapor pressure curve in relation to
            temperature
213       Double delta = (Double) ((4098 * (0.6018 * Math.exp((17.27 * t) / (t +
            273.3)))) / Math.pow((t + 237.2), 2));
214       // es - It is the saturated steam pressure (kPa)
215       Double es = (Double) (0.6108 * Math.exp((17.27 * t) / (t + 237.3)));
216       // ea - It is the current steam pressure (kPa)
217       Double ea = (Double) ((es * ru) / (100));
218       // Net radiation
219       Double rn = 0.0D;
```

```
220     // Estimation of net radiation
221     if (rs == null || rs == 0) { // if net radiation is null, the algorithm put
            an estimate value in this field
222       Double rns;
223       Double rnl;
224       Double krs = 0.16D;
225       Double rso;
226       Double ra;
227       Double dr;
228       Double phi;
229       Double omega_s;
230       Double sigma;
231       Double delta_lowercase;
232       Integer j;
233       Double x;
234       /** Calculation of rns **/
235       // j
236       Calendar cal = null;
237       cal = Calendar.getInstance();
238       cal.setTime(day);
239       j = cal.get(GregorianCalendar.DAY_OF_YEAR);
240       // phi
241       phi = (Double) ((lat * Math.PI) / 180);
242       // delta lowercase
243       delta_lowercase = (Double) (0.409 * Math.sin((2 * Math.PI) / (365) * j -
              1.39));
244       // x
245       x = (Double) (1 - Math.pow(Math.tan(phi), 2) * Math.pow(Math.tan(
              delta_lowercase), 2));
246       if (x <= 0) {
247         x = 0.00001D;
248       }
249       // omega s
250       omega_s = (Double) ((Math.PI / 2) - Math.atan((-Math.tan(phi) * Math.tan(
              delta_lowercase)) / (Math.pow(x, 0.5))));
251       // dr
252       dr = (Double) (1 + 0.033 * Math.cos((2 * Math.PI) / (365) * j));
253       // ra
254       ra = (Double) (118.08 / Math.PI * dr * (omega_s * Math.sin(phi) * Math.sin
              (delta_lowercase) + Math.cos(phi) * Math.cos(delta_lowercase) * Math.
              sin(omega_s)));
255       // rs
256       rs = (Double) (krs * ra * Math.sqrt(maxT - minT));
257       // rns
258       rns = (Double) (0.77 * rs);
259       /** Calculation of rnl **/
260       // rso
261       rso = (Double) ((0.75 + 2 * Math.pow(10, -5) * z) * ra);
262       // sigma
263       sigma = (Double) (4.903 * Math.pow(10, -9));
264       // rnl
265       rnl = (Double) (sigma * (((Math.pow((maxT + 273.16), 4)) + Math.pow((minT
              + 273.16), 4)) / 2) * (0.34 - 0.14 * Math.sqrt(ea)) * (1.35 * (rs /
              rso) - 0.35));
266       /** Calculation of rn **/
267       // rn
268       rn = rns - rnl;
```

```
269     } else {
270       Double rns;
271       Double rnl;
272       Double krs = 0.16D;
273       Double rso;
274       Double ra;
275       Double dr;
276       Double phi;
277       Double omega_s;
278       Double sigma;
279       Double delta_lowercase;
280       Integer j;
281       Double x;
282       /** Calculation of rns **/
283       // j
284       Calendar cal = null;
285       cal = Calendar.getInstance();
286       cal.setTime(day);
287       j = cal.get(GregorianCalendar.DAY_OF_YEAR);
288       // phi
289       phi = (Double) ((lat * Math.PI) / 180);
290       // delta lowercase
291       delta_lowercase = (Double) (0.409 * Math.sin((2 * Math.PI) / (365) * j -
              1.39));
292       // x
293       x = (Double) (1 - Math.pow(Math.tan(phi), 2) * Math.pow(Math.tan(
              delta_lowercase), 2));
294       if (x <= 0) {
295         x = 0.00001D;
296       }
297       // omega s
298       omega_s = (Double) ((Math.PI / 2) - Math.atan((-Math.tan(phi) * Math.tan(
              delta_lowercase)) / (Math.pow(x, 0.5))));
299       // dr
300       dr = (Double) (1 + 0.033 * Math.cos((2 * Math.PI) / (365) * j));
301       // ra
302       ra = (Double) (118.08 / Math.PI * dr * (omega_s * Math.sin(phi) * Math.sin
              (delta_lowercase) + Math.cos(phi) * Math.cos(delta_lowercase) * Math.
              sin(omega_s)));
303       // rns
304       rns = (Double) (0.77 * rs);
305       /** Calculation of rnl **/
306       // rso
307       rso = (Double) ((0.75 + 2 * Math.pow(10, -5) * z) * ra);
308       // sigma
309       sigma = (Double) (4.903 * Math.pow(10, -9));
310       // rnl
311       rnl = (Double) (sigma * (((Math.pow((maxT + 273.16), 4)) + Math.pow((minT
              + 273.16), 4)) / 2) * (0.34 - 0.14 * Math.sqrt(ea)) * (1.35 * (rs /
              rso) - 0.35));
312       /** Calculation of rn **/
313       // rn
314       rn = rns - rnl;
315     }
316     // main formula
317     dailyetopmf = (Double) ((0.408 * delta * (rn - g) + (lambda * 900 * u2 * (es
            - ea)) / (t + 273)) / (delta + lambda * (1 + 0.34 * u2)));
```

```
318       if (dailyetopmf == null || Double.isNaN(dailyetopmf)) {
319         dailyetopmf = 0.0D;
320       }
321       return dailyetopmf;
322     }
323   }
```