

## **Gestão do conhecimento científico usando uma arquitetura de grade dados para imagens FITS**

Maria Salete Marcon Gomes Vaz (UEPG e UFPR) salete@uepg.br  
Josiane Mariano Diniz Duszczak (UFPR) josiane\_m\_diniz@yahoo.com.br  
Lucélia de Souza (UNICENTRO e UFPR) lucelia@unicentro.br

**Resumo:** A necessidade de acesso, armazenamento e processamento de grande quantidade de dados, resulta em uma alta demanda de processamento, principalmente em astronomia, medicina, física, biologia e engenharia. Grades de dados são meios para prover e gerenciar recursos computacionais distribuídos para aplicações científicas, as quais demandam poder computacional ou utilizam equipamentos de uso específico na ciência. As grades de dados permitem a manipulação de grandes quantidades de dados e o compartilhamento coordenado e heterogêneo dos dados distribuídos. Já os metadados descrevem os dados, como meio de identificar e fornecer acesso e recuperação dos mesmos. Para manipulação dos metadados, é necessária uma padronização para favorecer a interoperabilidade dos dados. Quando não existe padronização dos metadados criados, o acesso e a recuperação dos dados e recursos da grade tornam-se tarefas difíceis devido à heterogeneidade dos metadados criados. O Formato FITS é utilizado para manipular, armazenar e transmitir imagens científicas. Tal formato é muito utilizado na Astronomia. Este artigo descreve a gestão do conhecimento científico usando uma arquitetura de grade dados para imagens FITS e a especificação de um padrão de metadados para gerenciamento desse tipo de imagens, estendido para o contexto das grades, como forma de promover a interoperabilidade, facilitando o acesso e a recuperação dessas imagens.

**Palavras chave:** Grade de Dados, Metadados, Padrões.

## **Scientific knowledge management using data grid architecture for FITS images**

**Abstract:** The need for access, storage and processing of data amounts large, implies in a high demand of processing, especially in astronomy, medicine, physics, biology and engineering. Data Grids are means of provide and manage distributed computing resources for scientific applications, which require computational power or the use of specific equipment's in science. Data grids allow handling large amounts of data and the coordinated sharing and heterogeneous of the distributed data. Metadata describe the data in order to identify and provide access and retrieval. To manipulation of the metadata, it is required standardization to ease the data interoperability. When the creation of metadata is not standardized, the access and the retrieval of the data from the grid become a difficult task, due to heterogeneity of these metadata. The FITS format is used to manipulate, store and transmit scientific images, being widely used in astronomy. This paper describes about the scientific knowledge management using data grid architecture to FITS images and the specification of a metadata standard for management of such images, extending it to the grids context, in order to promote interoperability, facilitating the access and the retrieval of these images.

**Keywords:** Data grid, Metadata, Standards

### **1. Introdução**

Cada vez mais existe a necessidade de acesso, armazenamento e processamento de grande quantidade de dados por entidades empresariais, acadêmicas e/ou governamentais. Em algumas áreas da ciência, a necessidade pela alta demanda de processamento se tornou obrigatória, como na astronomia, medicina, física, biologia e engenharia.

Neste cenário, o conceito de grades corresponde a um meio para prover e gerenciar recursos computacionais distribuídos para aplicações científicas, as quais demandam grande poder computacional ou utilizam equipamentos de uso específico na ciência (FOSTER, 2012).

As grades podem ser definidas como uma plataforma para execução de aplicações paralelas, em um ambiente heterogêneo e amplamente distribuído, não possuindo um controle central

(ROURE et al, 2005). Existe uma série de aplicações de ambientes de grades a serem desenvolvidas, assim como existem tecnologias que oferecem as funcionalidades necessárias.

Alguns dos benefícios da utilização da arquitetura de grade são: executar tarefas complexas com maior velocidade, possibilitar sistema unificado, permitir interface padronizada, executar/gerenciar cálculos e dados resultantes, transferir/trocar dados, e armazenar grande quantidade de dados.

De modo geral, as grades podem ser divididas nas categorias (KRAUTER et al., 2002) que seguem: (i) Grade computacional, utilizada para solucionar problemas computacionais complexos, tais como processamento de imagens médicas; (ii) Grade computacional oportunista, focando na utilização de ciclos computacionais ociosos; (iii) Grade de dados, a qual gerencia e distribui os dados, sendo o foco deste artigo; e (iv) Grade de serviços, que oferece serviços viabilizados pela integração de recursos na grade.

Uma grade de dados pode ser definida como um ambiente capaz de gerenciar os dados distribuídos, fornecendo suporte de acesso, sincronização e coordenação dos dados distribuídos, em locais remotos (BUYA et al., 2006). Nesses ambientes, usuários têm a possibilidade de acessar repositórios de dados e executar aplicações intensivas, aplicações que demandam grande poder computacional, proporcionando a análise de dados e o compartilhamento dos resultados gerados. Esses usuários podem visualizar os dados de locais diferentes.

Em grades de dados são manipuladas grandes quantidades de dados, permitindo o compartilhamento coordenado e heterogêneo dos dados distribuídos. Podem ser definidos metadados para descrever os dados, como meio de identificá-los para fornecer acesso e recuperação dos mesmos.

Para manipulação dos metadados, é necessária padronização para favorecer a interoperabilidade dos dados. Não havendo padronização, a heterogeneidade dos metadados criados torna as tarefas, como acesso e a recuperação dos dados e recursos da grade, não triviais.

No contexto deste artigo os dados gerenciados em grades são em Formato FITS – Sistema de Transporte de Imagens Flexíveis, usado para manipular, armazenar e transmitir imagens científicas. Esse formato é muito utilizado na Astronomia. Ao contrário de muitos formatos de imagem, é projetado especificamente para dados científicos e, portanto, inclui opções para descrever informações de dados espaciais, juntamente com os metadados de origem da imagem.

Na área de astronomia, muitos projetos de grades de dados estão em desenvolvimento, e não existe uma padronização específica para imagens do tipo FITS. Para tanto, uma arquitetura de grades de dados definida e o uso de padrões de metadados são necessários. Um padrão de metadados corresponde a um conjunto de informações (esquema de metadados) definido para atender um determinado contexto. Através da identificação de problemas no armazenamento e recuperação de informação por falta de padronização, vários esquemas são criados para atender diferentes propósitos (NISO, 2012).

Existem algumas questões críticas referentes aos metadados em ambientes de grades de dados (BUYA et al., 2006):

- Distribuição e Uniformidade de Acesso. Dados no ambiente de grades são distribuídos e gerenciados por organizações diferentes, onde os metadados associados aos dados seguem o mesmo princípio. Existe a necessidade de acesso uniforme aos repositórios de metadados

e ontologias, para oferecer visões integradas em grandes coleções de metadados distribuídos.

- Metadados devem ser acessíveis através de protocolos orientados a serviços, sendo necessário um modelo comum para gestão de metadados. No mínimo, é necessário manter a associação entre os recursos e metadados e apoiar as linguagens de consulta com os metadados específicos do modelo.
- Gerenciamento do ciclo de vida de metadados. Os metadados são dinâmicos e seu ciclo de vida pode variar em ordens de magnitude. É necessária a automação de todos os aspectos de gerenciamento de metadados, ou seja, formas de verificar mudanças de estado de metadados e propagá-los, usando uma infraestrutura de notificação.
- Controle de acesso uniforme e granular aos metadados. Duas questões são levantadas no controle de acesso. Deve ser cedida a permissão para acessar e manipular elementos de metadados, em diferentes níveis de granularidade, conforme o formato dos metadados, onde mecanismos de controle de acesso devem ser uniformes em todo repositório de metadados.

Atualmente, não existe a definição de um padrão para armazenamento de metadados de imagens FITS, em ambientes de grades de dados. Cada organização define um formato, conforme suas necessidades, afetando o desempenho na consulta e recuperação das imagens.

Os projetistas e/ou usuários de grades de dados criam suas próprias descrições de dados, de acordo com as necessidades de sua área de conhecimento. Na astronomia, para tratamento de imagens terrestres, existe um padrão mundialmente reconhecido denominado CSDGM – *Content Standard for Digital Geospatial Metadata* (CSFGM, 2012). Esse padrão não é adequado para o gerenciamento de imagens geoespaciais do tipo FITS, pois possui diversos metadados desnecessários, excessivos e/ou que não se aplicam para esse tipo de imagens, assim como não trata ambientes específicos de grade de dados.

Este artigo tem como objetivo apresentar uma abordagem arquitetural de grades de dados para imagens FITS e a especificação de um padrão de metadados para essas imagens científicas. A arquitetura tem como objetivo ajudar os cientistas a construir uma grade de dados e a especificação de um padrão de metadados facilita os processos de consulta e recuperação das imagens.

Este artigo está estruturado como segue. Além desta seção introdutória, existem mais 4 (quatro) seções. Na Seção 2 são descritos conceitos inerentes a grades de dados e metadados. Também, são apresentados conceitos de imagens FITS e Padrões de Metadados. Na Seção 3 é apresentado o padrão de metadados especificado para imagens FITS em grade de dados e a arquitetura de grade de dados para essas imagens. Na Seção 4 são abordados trabalhos relacionados e um comparativo entre eles. A Seção 5 apresenta as conclusões e perspectivas de pesquisas futuras.

## **2. Grade de Dados, Metadados e Padrões**

### **2.1. Grade de Dados**

Grade de dados pode ser definida como uma rede, onde o usuário se conecta para prestar e receber serviços computacionais. Corresponde a uma composição de infraestrutura de hardware e software, permitindo acesso a grandes capacidades computacionais, geograficamente distribuídas, de forma confiável, consistente, econômica e persistente (RAICU et al., 2010).

As grades utilizam recursos independentes e amplamente dispersos como plataforma de

execução de aplicações paralelas (FOSTER; KESSELMAN, 1999). Uma grade provê uma coordenação de recursos com o mínimo controle centralizado, baseado em padrões abertos, provendo uma qualidade não trivial de serviços.

Alguns benefícios proporcionados pelas grades são: (i) agregação de recursos; (ii) melhoria de qualidade e velocidade de serviços, acesso distribuído aos recursos, colaboração entre pesquisadores e; (iii) aproveitamento de recursos ociosos.

As organizações podem agregar recursos com toda a infraestrutura, não importando a localização geográfica. Na melhoria de qualidade e velocidade de serviços, as organizações podem melhorar os serviços disponibilizados devido à colaboração transparente dos recursos compartilhados.

O acesso distribuído aos recursos permite acesso e compartilhamento das bases de dados, de forma remota. A colaboração entre pesquisadores possibilita auxílio em projetos, por conta da habilidade no compartilhamento de experimentos, desde conceitos, estudos de casos e validações.

Existe aproveitamento dos ciclos de processamento ociosos disponíveis dos computadores, encontrados em várias localidades. Por exemplo, os computadores ociosos durante a noite em uma empresa em Tóquio, podem ser utilizados durante o dia para operações na América do Sul. Portanto, a tecnologia de grades possibilita agregar recursos computacionais, variados e dispersos, em um único supercomputador virtual, acelerando a execução de várias aplicações paralelas.

## **2.2. Metadados e Padrões**

Os metadados são definidos como dados que descrevem dados, ou, como a descrição detalhada das instâncias de dados, dos formatos, das características e dos valores dependentes dos mesmos (TANNENBAUM, 2002). Podem ser utilizados para descrever objetos ou, tornar pública sua existência.

Os metadados disponibilizam, descrevem, localizam e auxiliam na compreensão dos dados, transformando-os em conhecimento (VAZ, 2000). Ao ter conhecimento de quais dados são disponíveis, entender o seu contexto e onde são localizados, obtém-se informações precisas e melhores decisões podem ser tomadas (TANNENBAUM, 2002).

Ao serem identificados e registrados, os metadados são gerenciados como elementos dentro do repositório, sendo nomeados, os quais devem possuir um contexto, podendo conter informações sobre o domínio de negócio, áreas, sistemas, banco de dados, modelagem ou ambiente.

Os metadados podem ser classificados como (SENSO, PIÑERO, 2003): administrativos, descritivos, de preservação, técnicos, estruturais e pelo seu uso. Os metadados administrativos são utilizados na gestão de recursos de informação. Já os descritivos descrevem características de um documento, facilitando sua identificação, pesquisa e o gerenciamento das informações.

Os metadados de preservação são aqueles que salvaguardam as informações. Os técnicos estão relacionados ao funcionamento do sistema e do comportamento dos metadados. Os estruturais descrevem a forma como os objetos se interligam. Finalmente, os metadados de uso relacionam-se ao nível e ao tipo de uso dos recursos tecnológicos.

O conjunto de informações corresponde a um esquema de metadados definido para atender um determinado contexto. Através da identificação de problemas no armazenamento e recuperação de informação por falta de padronização, vários esquemas são criados para atender diferentes propósitos, chamados de padrões de metadados (NISO, 2012).

A utilização de padrões já foi vista como forma de limitação entre a comunidade de desenvolvedores. Atualmente, com o crescimento dos dados armazenados, padrões são vistos como aliados. O uso de padrões resulta em benefícios para a comunidade, tais como facilitar a atividade de análise, pois geralmente são amplamente documentados, e facilitam a comunicação entre os usuários, proporcionando uniformidade e integração entre soluções.

Dentre os padrões existentes para descrição de objetos, destacam-se:

- *Dublin Core Metadata Initiative* – DCMI (DCMI, 2012): define um grupo de quinze atributos que pode ser utilizado para descrever seus próprios recursos na *Web*, destacando-se pela simplicidade, interoperabilidade semântica, consenso internacional, extensibilidade e modularidade. Esse padrão pode ser utilizado nas mais diversas áreas de conhecimento, devido a utilizar atributos genéricos.
- *Metadata Encoding and Transmission Standard* – METS (METS, 2012): é um padrão implementado em XML para a codificação de metadados descritivos, administrativos e estruturais, utilizados para a gestão e a troca de objetos de repositórios de bibliotecas de objetos digitais. Tais bibliotecas requerem a manutenção de vários tipos de metadados estruturados. Esse padrão possui sete seções, em cada seção possui um grupo de atributos, sendo: Cabeçalho METS, Metadados Descritivos, Metadados Administrativos, Seção de Arquivos, Mapa Estrutural, Ligações Estruturais e Comportamento. Esse padrão tem como vantagem, ser coerente e flexível para descrição de objetos de bibliotecas digitais.
- *Movie Picture Experts Group - MPEG 7* (MPEG, 2012): utilizado para descrever dados de áudio e vídeo. A interface de descrição de conteúdo multimídia foi desenvolvida para prover um *template* utilizado em repositórios de dados para recuperação automatizada em aplicações. Esse padrão é adequado para a descrição das informações estruturais e semânticas de conteúdos multimídia.
- *Learning Object Metadata* – LOM (DUVAL et al., 2012): utilizado para descrição de objetos de aprendizagem, podendo ser qualquer entidade, digital ou não, usada, reusada ou referenciada no aprendizado, em meio tecnológico. O Padrão LOM usa uma abordagem estruturada para criação de metadados, utilizando nove categorias: Geral, Ciclo de vida, Meta-metadados, Técnica, Educacional, Direitos, Elação, Anotação e Classificação. Apesar de criar relacionamentos hierárquicos complexos, facilitando a descoberta de recursos, é um padrão de difícil utilização por usuários iniciantes, podendo resultar em registros de metadados incompletos ou insuficientes para pesquisa/recuperação de recursos com qualidade.
- *Content Standard for Digital Geospatial Metadata* – CSDGM (FGDC, 1998): possui um total de 245 elementos de metadados, divididos em sete grupos (descritos na próxima seção), requerendo uma excessiva quantidade de informações. É de aplicação questionável devido à exigência do preenchimento de um conjunto de formulários exaustivos.
- *MetaFits* (NOGUEIRA et al., 2010): padrão de metadados para documentação de imagens geoespaciais do formato FITS, sendo formado por 30 atributos, distribuídos em cinco categorias. Esse padrão é descrito em mais detalhes na próxima seção.

O uso de padrões de metadados permite facilitar a decisão de quais metadados devem ser coletados e mantidos, visto que os metadados podem ter uma variedade de formas. Além do fato de que diferentes comunidades podem usar o mesmo padrão e não propor diferentes tipos de metadados e/ou adotar vocabulários diferentes, facilitando comunicação e a interoperabilidade.

O Padrão de Metadados CSDGM é utilizado para a documentação de imagens geoespaciais digitais, ou seja, imagens obtidas da superfície terrestre, a partir de satélites. As imagens FITS são obtidas a partir de telescópios localizados na superfície terrestre que fotografam, por exemplo, galáxias, constelações e superfícies planetárias (Figura 1).



Nebulosa Trífida

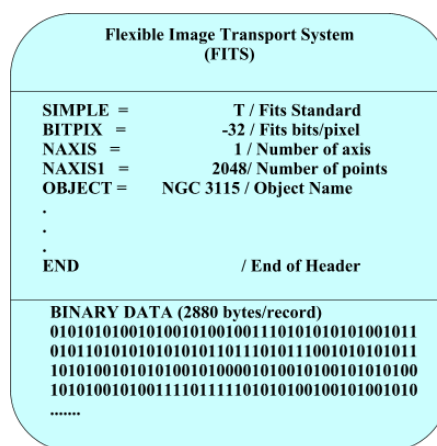


Galáxia

**Figura 1: Imagens no formato FITS fotografadas pelo Telescópio Espacial Hubble**

O formato de arquivo FITS (IAU, 2012) é comumente utilizado em astronomia, sendo um formato utilizado, principalmente, por instituições de pesquisa espacial, para armazenamento e troca de imagens astronômicas. Ele oferece parâmetros que podem estar associados às imagens (HANISCH, 2001).

Um arquivo FITS é composto pela imagem e por um cabeçalho associado à mesma. O arquivo utiliza o formato binário para o armazenamento das imagens e o formato ASCII para o armazenamento do cabeçalho (Figura 2), permitindo aos pesquisadores e máquinas o reconhecimento das informações contidas nesse cabeçalho.



**Figura 2: Esquematização de um arquivo no formato FITS.**

Cada cabeçalho é constituído por vários registros, os quais são compostos por um valor precedido por uma palavra-chave correspondente. Esses pares palavra-chave/valor fornecem metadados como nome, data, tamanho, origem, instrumento, coordenadas, comentários e história da imagem ao qual estão associados.

Em um cabeçalho, são seis as palavras-chaves obrigatórias (WELLS et al., 1981): 1) SIMPLE (tipo lógico) especifica se o arquivo está de acordo com as normas FITS; 2) BITPIX (tipo inteiro) especifica o número de bits utilizado para representar cada valor de *pixel*; 3) NAXIS (tipo inteiro) especifica o número de dimensões (eixos de coordenadas) da imagem; 4)

NAXISn (tipo inteiro) informa o número de *pixels* utilizados na dimensão n; 5) OBJECT (tipo *string*) identifica o nome do objeto; 6) END indica o fim do arquivo.

O Metafits foi criado para a documentação das imagens FITS baseado nos grupos de metadados do Padrão CSDGM, uma vez que este tem sido a base para o desenvolvimento dos principais padrões de metadados existentes, porém o Metafits tem como maior benefício uma quantidade menor de atributos obrigatórios, focando em imagens obtidas através de telescópios localizados na superfície terrestre.

Um padrão de metadados precisa ser claro, compreensível, consistente, completo, flexível e, principalmente, fácil de usar. Um número mínimo de metadados pode garantir uma estrutura de metadados homogênea, confiável e robusta (WEBER et al. 1990).

A maioria dos padrões utiliza uma quantidade excessiva de metadados de caráter administrativo, quando aspectos de qualidade teriam maior relevância no uso e implementação de metadados. Como no Padrão CSDGM, as informações incluídas no MetaFits (NOGUEIRA et al., 2010), para a documentação de imagens FITS, foram selecionadas com base em quatro características necessárias (WEBER et al. 1990): (i) Disponibilidade, metadados para descrever o conjunto de dados para localização geográfica; (ii) Adequação para uso, determinando se um conjunto de dados preenche uma necessidade específica; (iii) Acesso, estabelecendo os dados necessários para introdução de um conjunto de dados; e (iv) Transferência, definindo os dados necessários para processar e utilizar um conjunto de dados.

Baseado nisso, foi desenvolvido um padrão com as características de metadados inerentes às imagens FITS e adaptado a partir dos grupos de metadados do CSDGM. As palavras-chave do cabeçalho foram empregadas como metadados do MetaFits e novos metadados foram adicionados, objetivando reduzir a utilização dos metadados administrativos e focando nos metadados que referenciam os aspectos de qualidade.

Para documentação de imagens FITS, cinco grupos de metadados foram definidos baseados nos grupos utilizados pelo Padrão CSDGM:

- *Identification*: compreende os metadados responsáveis pela caracterização básica das imagens, bem como pela identificação da mesma;
- *Quality*: agrupa os metadados inerentes à qualidade da imagem, como sua definição e clareza.
- *Coordinate System*: abrange os metadados responsáveis pela localização espacial do objeto presente na imagem.
- *Source*: contém informações a respeito das pessoas, entidades e/ou órgãos de pesquisa responsáveis pela geração da imagem FITS.
- *Reference*: contém os metadados responsáveis por informar quando os metadados foram gerados e quem foi responsável pela geração dos mesmos.

O MetaFits tem como vantagem a garantia de uma estrutura de dados sem o emprego excessivo de metadados administrativos, comparado com a maioria dos padrões de metadados para a documentação de imagens geoespaciais.

A desvantagem da sua utilização é a pequena quantidade de dados cadastrados e disponíveis para manipulação. Essa desvantagem tende a ser superada, de médio e longo prazo, através da utilização desse padrão e do aumento da quantidade de dados cadastrados segundo os critérios estabelecidos. Padrões utilizados há algum tempo, como o CSDGM possuem grandes bases de dados catalogadas de acordo com seus critérios.

### 2.3. Grades de Dados e Metadados

Metadados em ambientes de grades auxiliam na recuperação, localização, acesso e gerência de dados. Algumas informações que os metadados descrevem são: proveniência dos dados, informações físicas e autoridades de acesso sobre os dados.

A proveniência dos dados é inerente a como os itens de dados são criados ou transformados, e por quais instrumentos científicos. Os metadados de informações físicas fornecem informações de tamanho e localização. Os metadados inerentes a autoridades de acesso sobre os dados descrevem as informações sobre os proprietários e leitores de dados.

Os metadados para grades de dados incluem três aspectos: informações de sistemas, de réplicas e de aplicativos (MAGOULÈS et al.,2009). As informações de sistemas abrangem informações estruturais sobre as grades de dados. Por exemplo, condições de serviços sobre a *Internet*, capacidade de armazenamento dos dispositivos de armazenamento, *status* de ociosidade do computador e políticas de uso.

As informações de réplicas descrevem o mapeamento entre arquivos lógicos e cópias físicas. Quanto às informações de aplicativos, estas descrevem atributos definidos pela comunidade. Por exemplo, conteúdo dos dados e as informações semânticas sobre os dados.

O processo de publicação necessita que os serviços de metadados possam combinar determinadas informações dos metadados com o conjunto de dados armazenados, para que após isso os dados possam ser utilizados pelo serviço de descoberta dos dados.

A descoberta de dados é um processo de identificação dinâmica, que identifica o conjunto de dados pertinentes e sua localização através de informações de atributos de consultas publicadas pelo serviço de metadados, ou informações como especificação da estrutura interna, membros, proveniência ou propriedades físicas, como tamanho e caminho de acesso.

Após a descoberta dos requisitos dos itens de dados, é possível ao usuário acessar os dados originais ou réplicas. Os serviços de descoberta são capazes de mostrar o conteúdo dos recursos de dados de uma forma estática.

### 3. Arquitetura de Grade de Dados Para Imagens FITS

Nesta seção é apresentada a arquitetura de grade de dados para imagens FITS, com o Padrão de Metadados MetafitsGrid. A arquitetura é dividida em 4 camadas (Figura 3): Camadas de Metadados, de Aplicação, Operacional e de Hardware.

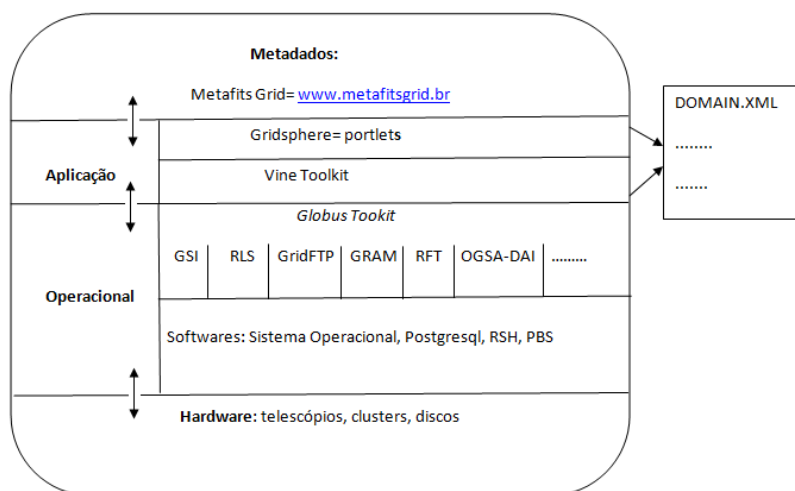


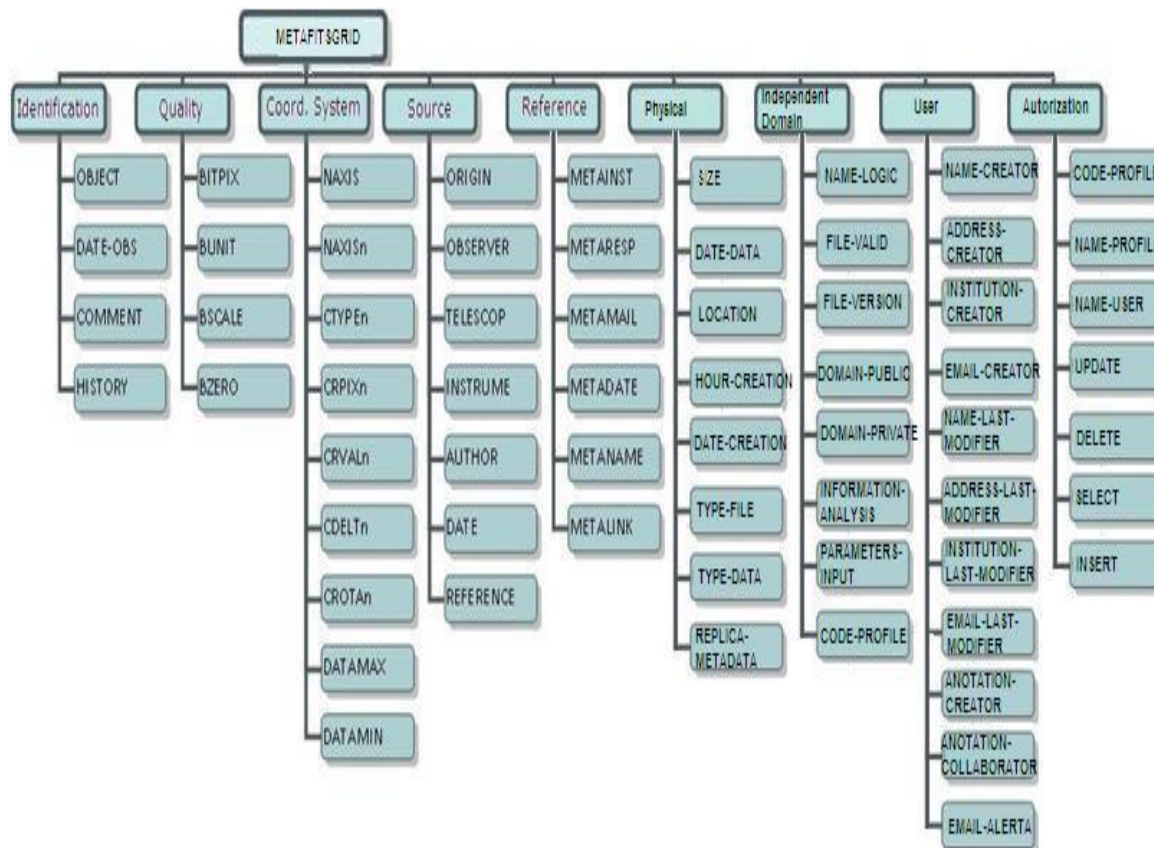
Figura 3: Visão geral – Arquitetura de Grade de Dados para Imagens FITS



### Camada de Metadados

A camada de Metadados permite aos usuários entenderem o significado dos dados e o contexto da aplicação. Permite ao administrador mapear os dados de acordo com um esquema de classificação, objetivando criar uma visão das informações da aplicação.

Na Camada de Metadados foi o definido neste artigo o Padrão MetafitsGrid para descrever imagens FITS, em ambientes de Grades, o qual estende o Padrão MetaFits. Os metadados foram divididos em 4 grupos, sendo definidos 34 metadados (Figura 4).



**Figura 4: Organograma do Padrão MetafitsGrid**

**Metadados físicos:** incluem informações sobre características físicas dos dados. No caso do *Type-Data*, cada sistema gerenciador de dados tem sua própria definição para tipo de dados e, estas definições não são únicas. Como exemplo, em *MySQL* e *Oracle*, existem dois diferentes nomes para o mesmo tipo de dados, *char* e *varchar*.

O tipo de metadados *Replica-Metadata* registra a ligação entre o arquivo lógico de dados e a réplica de um ou mais arquivos físicos, podendo estar armazenados em sistemas de arquivo ou base de dados diferentes. Este tipo de metadados é frequentemente usado pelo sistema gerenciador de metadados.

Os metadados de domínio independente são aqueles que se aplicam a qualquer item de dados. Nesta categoria, os metadados de proveniência estão especificados, os quais descrevem transformações ocorridas na grade.

O *File-Valid* é um tipo de metadados válido e indica se um item de dados é atualmente válido, permitindo rapidamente invalidar arquivos lógicos. Se existirem várias versões especificadas, tanto o nome de arquivo lógico e o número de versão devem ser fornecidos para permitir que o gerenciador de metadados identifique o item de dados desejado. Isso é determinado pelo tipo de metadados *File-Version*.

Os metadados inerentes aos usuários incluem informações referentes aos mesmos. Já os metadados de autorização são aqueles que especificam os privilégios de acesso em arquivos lógicos.

O principal objetivo da camada de metadados é possuir padrão para descrição de imagens FITS na grade, facilitando o acesso aos dados, através de consultas.

### **Camada de Aplicação**

A Camada de Aplicação tem as funcionalidades necessárias à adaptação dos processos de aplicação ao ambiente de comunicação. É estruturada para permitir a flexibilidade das funções e de forma, para se determinar os requisitos de comunicação de cada aplicação distribuída.

Esta camada apresenta uma interface amigável para o usuário final da grade de dados. Pode ser instanciada através de dois componentes, o *Gridsphere* e *Vine Toolkit*. O *Gridsphere* se comunica diretamente com a camada de Metadados para consultar as descrições realizadas através do padrão de metadados proposto e o *Vine Toolkit* se comunica diretamente com o *Middleware Globus*. A comunicação entre o *middleware Globus* e o *framework Gridsphere* é feita através do *Vine Toolkit*.

*Vine* é um sistema modular e uma biblioteca Java extensível, oferecendo aos desenvolvedores APIs para aplicações de grade. Já o *Gridsphere* permite a criação de um *framework* para o desenvolvimento de portais de grades baseado no conceito de *portlets*. Desenvolvedores têm a possibilidade de criar aplicações a serem executadas e administradas dentro do *Gridsphere*.

### **Camada Operacional**

A camada operacional é responsável pelos softwares instalados para o funcionamento da grade. Essa camada corresponde ao núcleo da grade de dados, com configuração de aspectos de segurança, escalonadores e gerenciadores de réplicas. Esta camada faz a comunicação com os recursos de hardware e fornece informações para a camada de aplicação (permissões de acesso, por exemplo).

### **Camada de Hardware**

A camada de hardware é composta pelos computadores pessoais, *clusters* e quaisquer outros recursos de hardware conectados à grade de dados.

## **4. Trabalhos Relacionados**

Nesta seção são descritos trabalhos relacionados AstroGrid-D (ENKE et al., 2010), BD-Grid (XIE et al., 2007) e GridMACS (CHIA et al., 2010), onde são enfatizados grades de dados e padrões de metadados.

O AstroGrid constitui uma grade de dados astronômica, desenvolvido na Alemanha, que tem como objetivo compartilhamento de experimentos e maior rapidez de execução, utilizando o *middleware Globus*. Desenvolvedores do AstroGrid criaram 4 (quatro) categorias de metadados: metadados de recursos, metadados de estado de atividade, metadados de aplicação e metadados científicos.

O BD-Grid trata de grade de dados para analisar o conjunto de dados biológicos, baseado no *Middleware GD-Grid*, possuindo um módulo de categorias de metadados, as quais são mapeadas para um catálogo do BD-Grid. Através do Portal do BD-Grid, o usuário pode publicar seus metadados ou encontrar seus dados. Porém, no BD-Grid não existe um mecanismo para validar metadados incluídos, ou seja, não se pode garantir que os metadados inseridos sejam confiáveis.

O GRIDMACS trata grade de dados biológicos, focada em simulações de moléculas de proteínas. Apresenta uma interface amigável para usuários finais. Utiliza ferramentas para construção da grade de dados: Globus Toolkit 4, Gridsphere e Vine Toolkit. Não há abordagem de conceitos de metadados e o projeto teve problemas para implementação da grade, devido à incompatibilidade de versões. Há emprego de tecnologias de grade, provendo uma plataforma para facilitar simulações, aumentando a disponibilidade e a usabilidade.

A arquitetura grade de dados proposta neste trabalho para imagens FITS foca na área astronômica. Sua arquitetura é dividida em camadas: Hardware, Operacional, Aplicação e Metadados. Define um padrão de metadados para imagens FITS, em ambientes de grades de dados, com o objetivo de facilitar o acesso aos dados cadastrados.

Os metadados foram divididos nas categorias Identificação, Qualidade, Sistema de Coordenadas, Fonte, Referências, Físicos, Independente de Domínio e Usuários. Utiliza o *middleware* Globus Toolkit 4, Vine Toolkit e Gridsphere.

## 5. Conclusões e Perspectivas de Trabalhos Futuro

Grades de dados permitem o compartilhamento de dados e experimentos, necessidade dos cientistas que estão localizados em lugares geograficamente distantes. Muitos experimentos necessitam de grande poder computacional e armazenamento. Construir e utilizar uma grade de dados não é uma tarefa trivial.

Este artigo apresentou uma abordagem arquitetural de Grade de dados para imagens FITS, descrevendo as camadas de um ambiente de grade e seus componentes. É apresentada a especificação de um padrão de metadados para gerenciamento desse tipo de imagens, estendido para o contexto das grades, como forma de promover a interoperabilidade, facilitando o acesso e a recuperação dessas imagens.

Também alcançou o objetivo de facilitar a criação de um ambiente de grade de dados para imagens FITS, apresentando componentes que são essenciais e alguns softwares que podem tornar o ambiente de grade mais amigável, como o uso de gerenciadores de portais, tais como Gridsphere e Vine.

Porém, a arquitetura proposta pode ser utilizada para construir arquiteturas de grades de dados em diversas áreas de conhecimento. O padrão de metadados proposto facilita a recuperação e a consulta aos dados de imagens FITS, onde o padrão pode ser reutilizado para ambientes de grades de dados de outras áreas de conhecimento, alterando apenas os metadados da área de conhecimento.

Como perspectivas de trabalhos futuros, será realizada a comunicação entre o Globus Toolkit, Vine e Gridsphere, realizar testes com o padrão de metadados proposto, armazenando e consultando imagens FITS e automatizando o processo de inclusão dos metadados.

## Referências

**BUYA, R; VENUGOPAL, S; RAMAMOCHANARAO, K.** *A Taxonomy of Data Grids for Distributed Data Sharing, Management, and Processing.* University of Melbourne, Australia. ACM Computing Surveys, Vol. 38. Article 3, 2006.

**CHIA, Elizabeth; SHAMSIR, Shahir Mohd; HUSSEIN, Azura Zeti; HASHIM, Siti Zaitom Mohd.** *GridMACS Portal: A Grid Web Portal for Molecular Dynamics Simulation using GROMACS.* 2010 Fourth Asia International Conference on Mathematical/Analytical Modelling and Computer Simulation, 2010.

**CSDGM** – *Content Standard for Digital Geospatial Metadata.* Disponível em: <http://www.fgdc.gov/metadata/csdgm/>. Acesso em: Ago/2012.

**DCMI** - *Dublin Core Metadata Initiative.* Disponível em: <http://dublincore.org/> Acesso em: Ago/2012.

- DUVAL, E., HODGINS, W., SUTTON, S., WEIBEL S.** *Metadata Principles and Practicalities*. D-Lib Magazine 8(4). Disponível em: <http://www.dlib.org/dlib/april02/weibel/04weibel.html>. Acesso em: Ago/2012.
- ENKE, H.; STEINMETZ, M.; ADORF, H. ; BECK-RATZKA A.; BREITLING, F.; BRUSEMEISTER, T.; CARLSON, A.; ENSSLIN, T.; HOGQVIST, M.; NICKELT, I.; RADKE, T.; REINEFELD, A.; REISER, A.; SCHOLL, T.; SPURZEM, R.; STEINACKER, J.; VOGES, W.; WAMBSGANB, J.; WHITE, Steve.** *AstroGrid-D: Grid Technology for Astronomical Science*, 2010.
- FGDC - Federal Geographic Data Committee.** *Content Standard for Digital Geospatial Metadata*. User Guide, 1998.
- FOSTER, I.** *What is the Grid? A Three Point*. Disponível em: <http://dlib.cs.odu.edu/WhatIsTheGrid.pdf>. Acesso em: Ago/2012.
- FOSTER, I; KESSELMAN, C.** *The Grid: Blueprint for a Future Computing Infrastructure*, 1999.
- HANISCH R. J.; FARRIS A; GREISEN E. W.; PENCE W. D.; SCHLESINGER B. M.; Teuben, P. J.; THOMPSON R. W.; WARNOCK III A.** *Definition of the Flexible Image Transport System (FITS)*. NASA/ Science Office of Standards and Technology, 2001.
- IAU.** *International Astronomical Union*. Disponível em: <http://www.iau.org/> Acesso em: Ago, 2012.
- KRAUTER, K; BUYYA, R; MAHESWARAN, M.** *A Taxonomy and Survey of Grid Resource Management Systems for Distributed Computing*. Software: Practice&Experience,32(2):135–164, 2002.
- MAGOULÈS, F; PAN, J; TAN, An K; KUMAR, A.** *Introduction to Grid Computing*. Taylor & Francis Group, LLC, 2009.
- METS-Metadata Encoding and Transmission Standard (METS).** Disponível em <http://www.loc.gov/standards/mets/>. Acesso em: Ago/2012.
- MPEG-Moving Picture Experts Group (MPEG) ISO/IEC JTC1 SC29 WG11.** Disponível em <http://www.chiariglione.org/mpeg/>. Acesso em Ago/2012.
- NISO.** *Understanding Metadata*. Disponível em: <http://www.niso.org/publications/press/UnderstandingMetadata.pdf>. Acesso em: Ago/2012.
- NOGUEIRA, Eduardo Dimas Andrino ; VAZ, M. S. M. G. ; SOUZA, Lucélia de.** *Um Padrão de Metadados para Descrição de Imagens Astronômicas do Tipo FITS*. Revista Ciências Exatas e Naturais, v. 12, p. 55-72, 2010.
- RAICU, I.; FOSTER, I.T.; WILDE, M.; ZHANG, Z.; ISKRA, K., BECKMAN, P.H.; ZHAO, Y.; SZALAY, A.S., CHOUDHARY, A.N.; LITTLE, P.** *Middleware support for many-task computing*. Cluster Computing(2010) 291-314, 2010.
- ROURE, D.; JENNINGS, N. R.; SHADBOLT, N. R.** *The Semantic Grid: Present, past and future*. Proceedings of IEEE 2005. Pages 669-681, 2005.
- SENSO, J. A.; PIÑERO, A. R.** *El concepto de metadato*. Algo más que descripción de recursos electrónicos. Ciência da Informação, Brasília, v. 32, n. 2, maio/2011, 2003.
- TANNENBAUM, A.** *Metadata Solutions: Using Metamodels, Repositories, XML and Enterprise Portals to Generate Information on Demand*. Addison Wesley, 2002.
- VAZ, M. S. M. G.** *MetaMídia – Um Modelo de Metadados na Indexação e Recuperação de Objeto Multimídia*. Tese (Doutorado em Ciência da Computação) – Universidade Federal de Pernambuco, 2000.
- VINE TOOLKIT 1.3.2.** Disponível em: [http://vinetoolkit.org/software\\_releases](http://vinetoolkit.org/software_releases), Acesso em: fev/2011.
- XIE, Jiang; ZHANG, Wu; MEI, Jian.** *A Data Grid System oriented Biologic Data*. IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology – Workshops, 2007.
- WELLS D. C.; GREISEN E. W.; HAETEM R. H.** *FITS: A Flexible Image Transport System*. Astronomy & Astrophysics Supplement Series. Páginas 363-370, 1981.
- WEBER E.; ANZOLCH R.; LISBOA F. J.; COSTA A. C.; IOCHPE C.** *Qualidade de Dados Geoespaciais*. Relatório de Pesquisa. Universidade Federal do Rio Grande do Sul, 1990.